Employee Retention Analysis

- Bijender Singh

Agenda

- **Introduction**
- Brief overview of the problem statement
- Importance of the problem and its relevance in real-world applications
- Objective of the assignment and what it aims to solve
- **Assumptions Made**
- List any assumptions made during the analysis and solution development
- Mention how these assumptions might impact the results

Problem Statement

- **Objective:** A mid-sized technology company aims to predict employee retention to improve workforce stability.
- **Goal:** Build a logistic regression model to predict the likelihood of employees staying.
- **Impact:** Provide actionable insights for HR to strengthen retention strategies and foster a more committed and loyal workforce.

Handling Missing Values

Characterizing Missing Values

Understanding the types and causes of missing values—whether they are missing completely at random or due to specific patterns—helps determine the best strategies for handling them.

Mean Imputation Technique

Using the mean for imputation can smooth out missing information, yet it assumes no significant skewness in the underlying data, a consideration to be mindful of in analysis.

Balanced data identification

Effective strategies must aim to minimize bias introduced during the imputation process to uphold the integrity of analytical outcomes and support fair interpretations.

Exploratory Data Analysis (EDA)

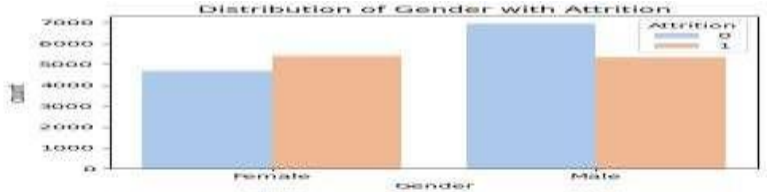
Visualizing Data Distributions

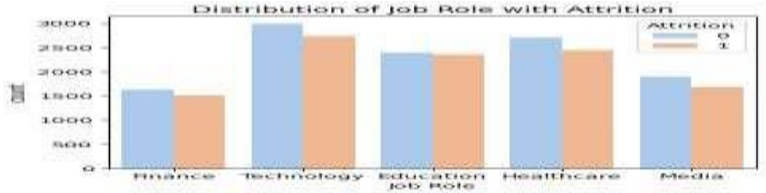
Interpreting Correlation Coefficients

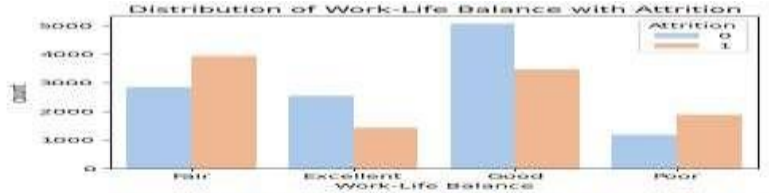
Applications of Correlation in EDA

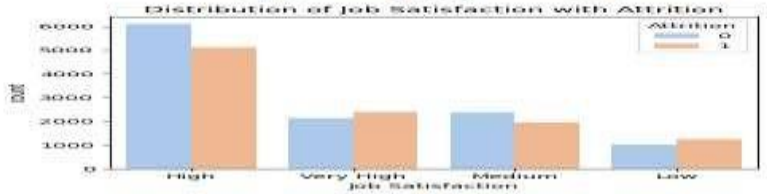
Limitations of Correlation Analysis

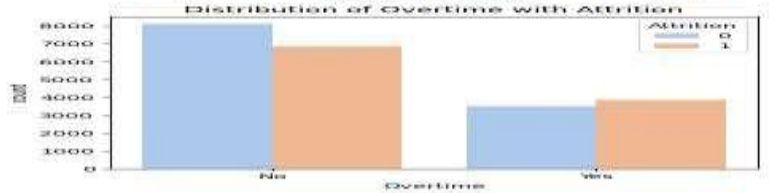
Perform bivariate analysis

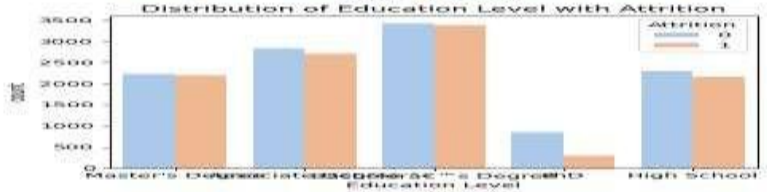


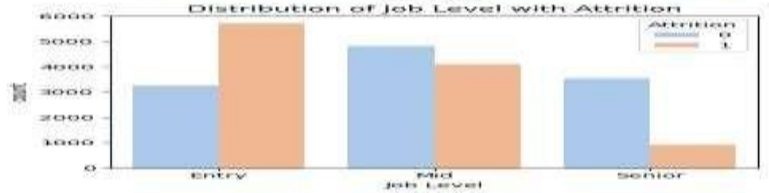


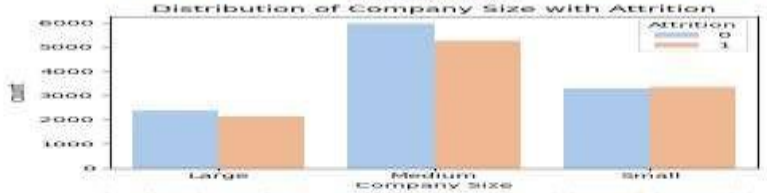


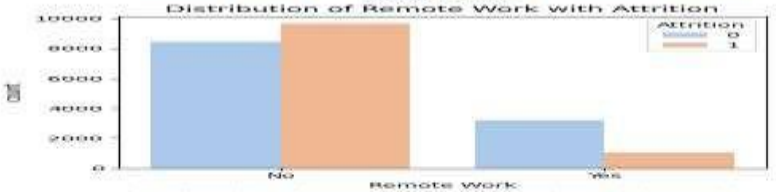


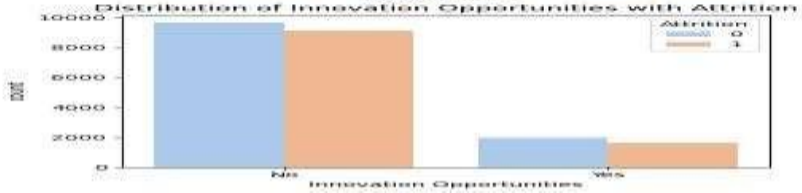


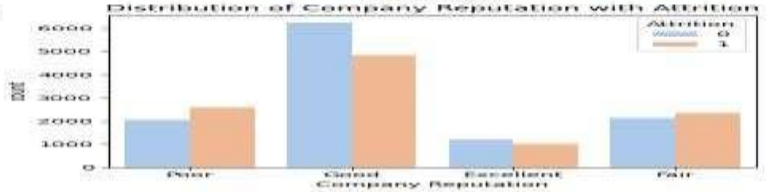


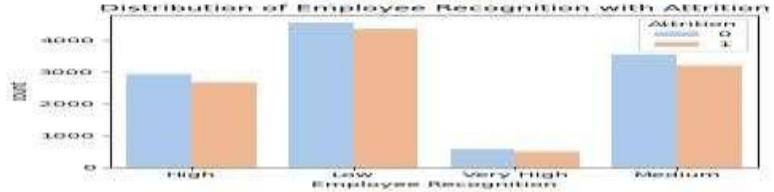




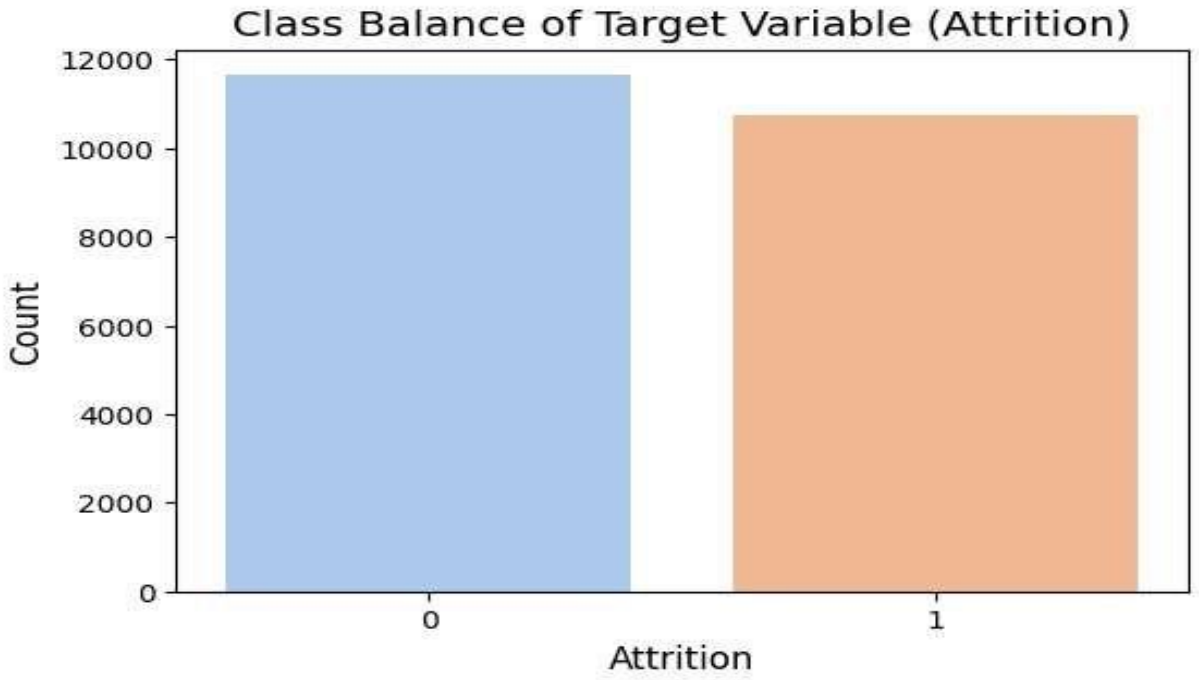




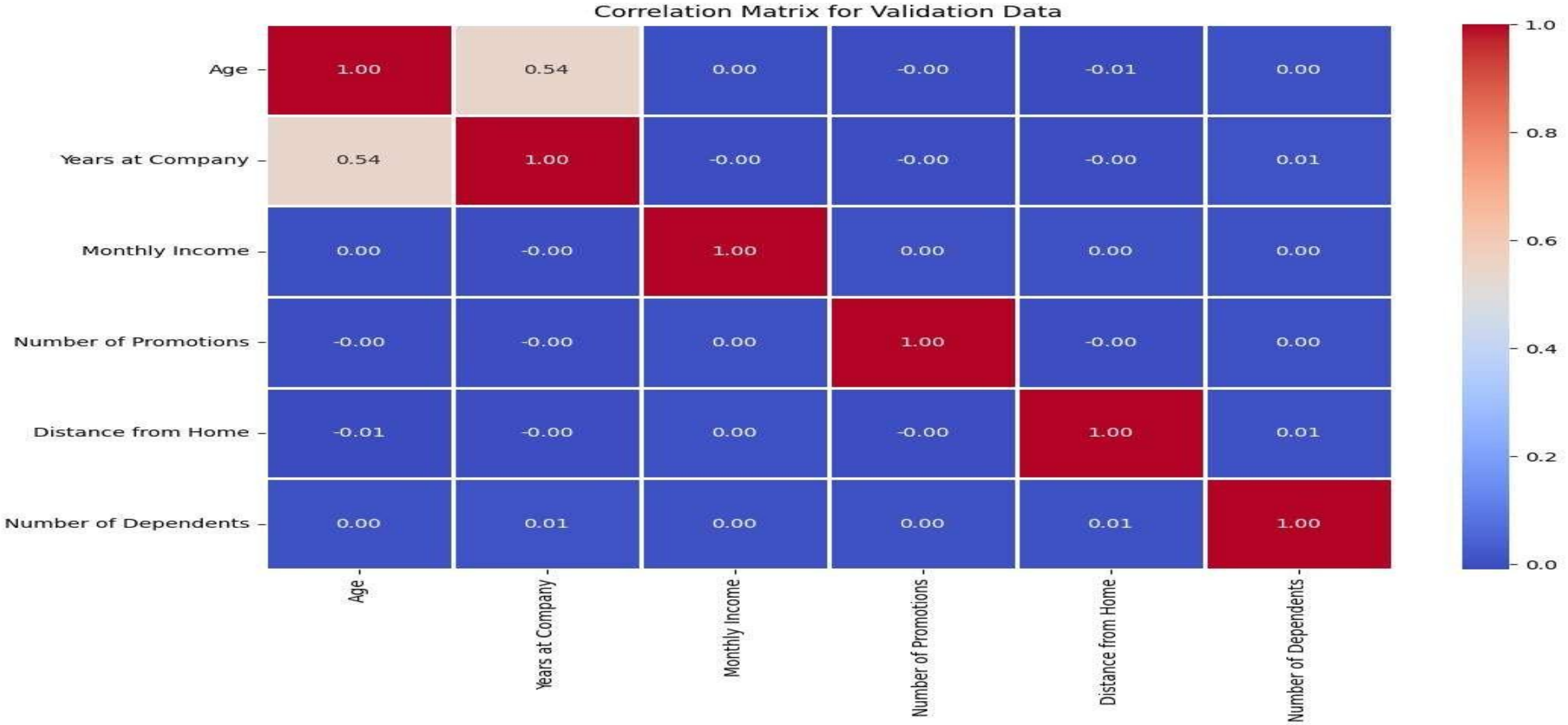




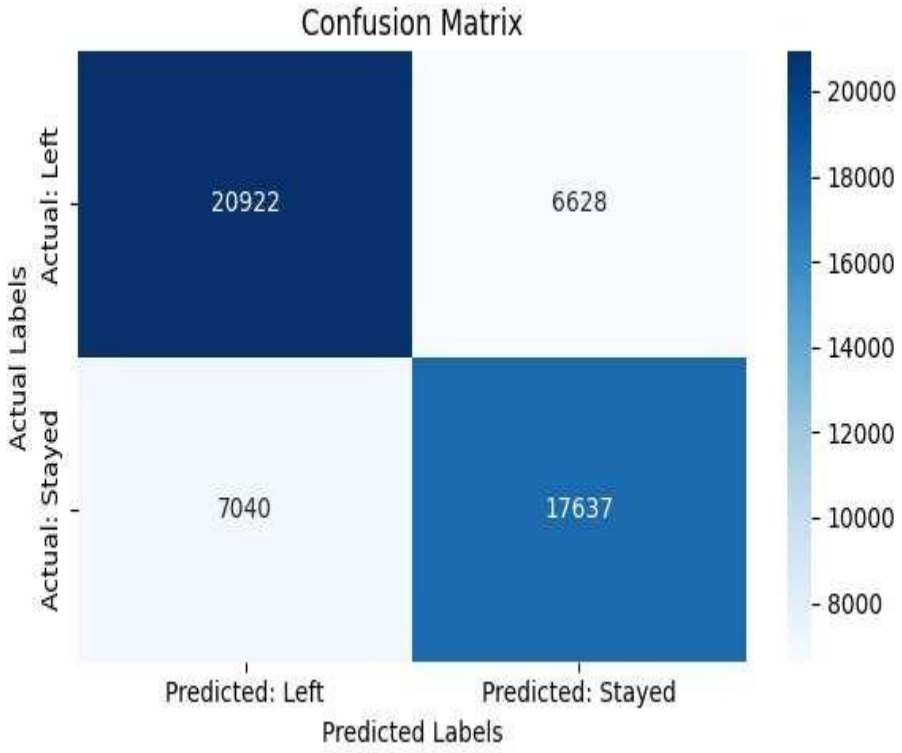
Check class balance



- **Balanced Dataset:** Indicates that the number of instances in each is nearly equal.
- **Fair Model Training:** Prevents model bias toward the majority class, leading to more reliable and unbiased predictions.
- **Improved Metric Reliability:** Metrics like accuracy, precision, and recall become more meaningful when classes are balanced.
- **No Need for Resampling:** Balanced data reduces the need for techniques like oversampling or undersampling.

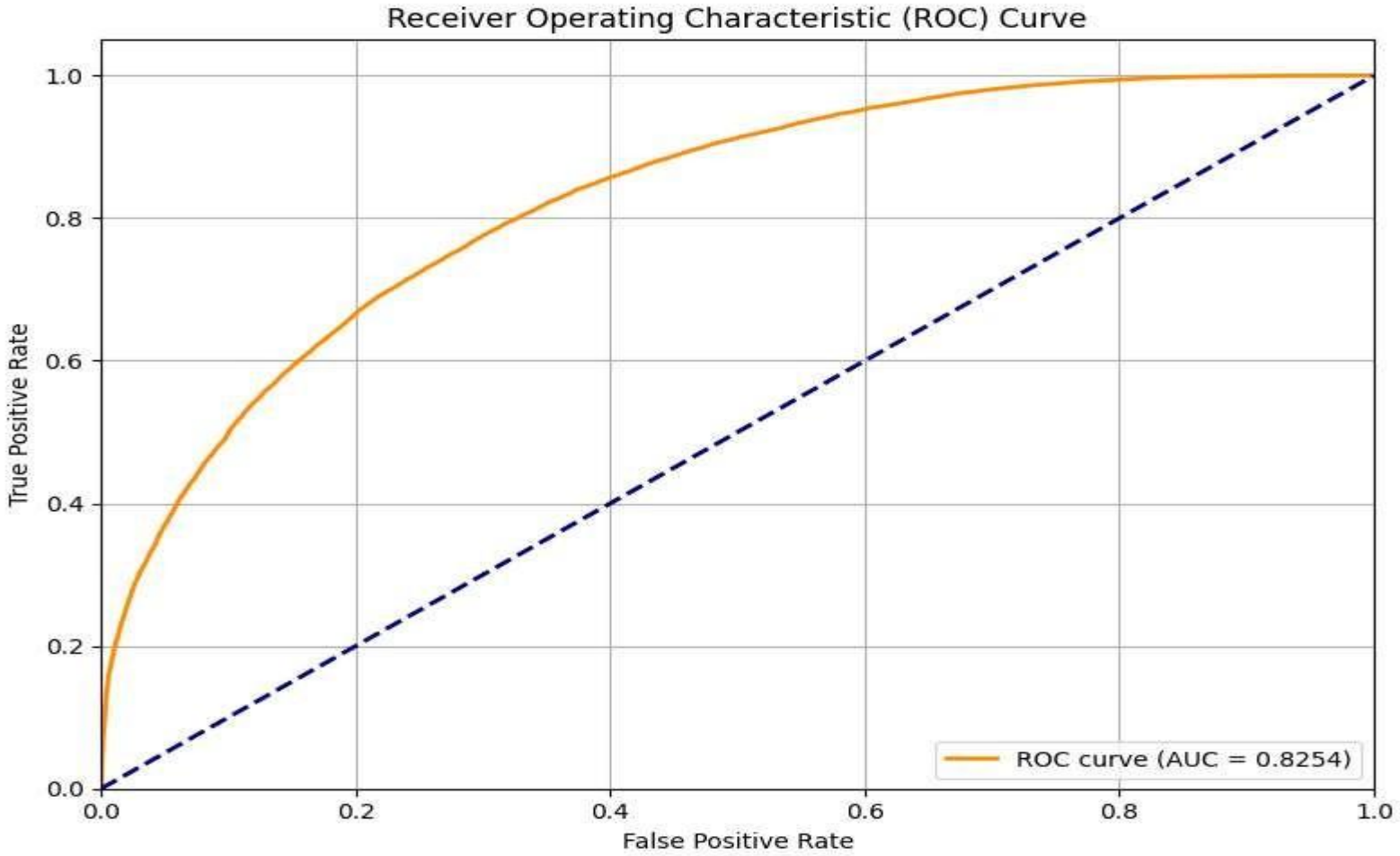


key insights: `Confusion Matrix Analysis:`



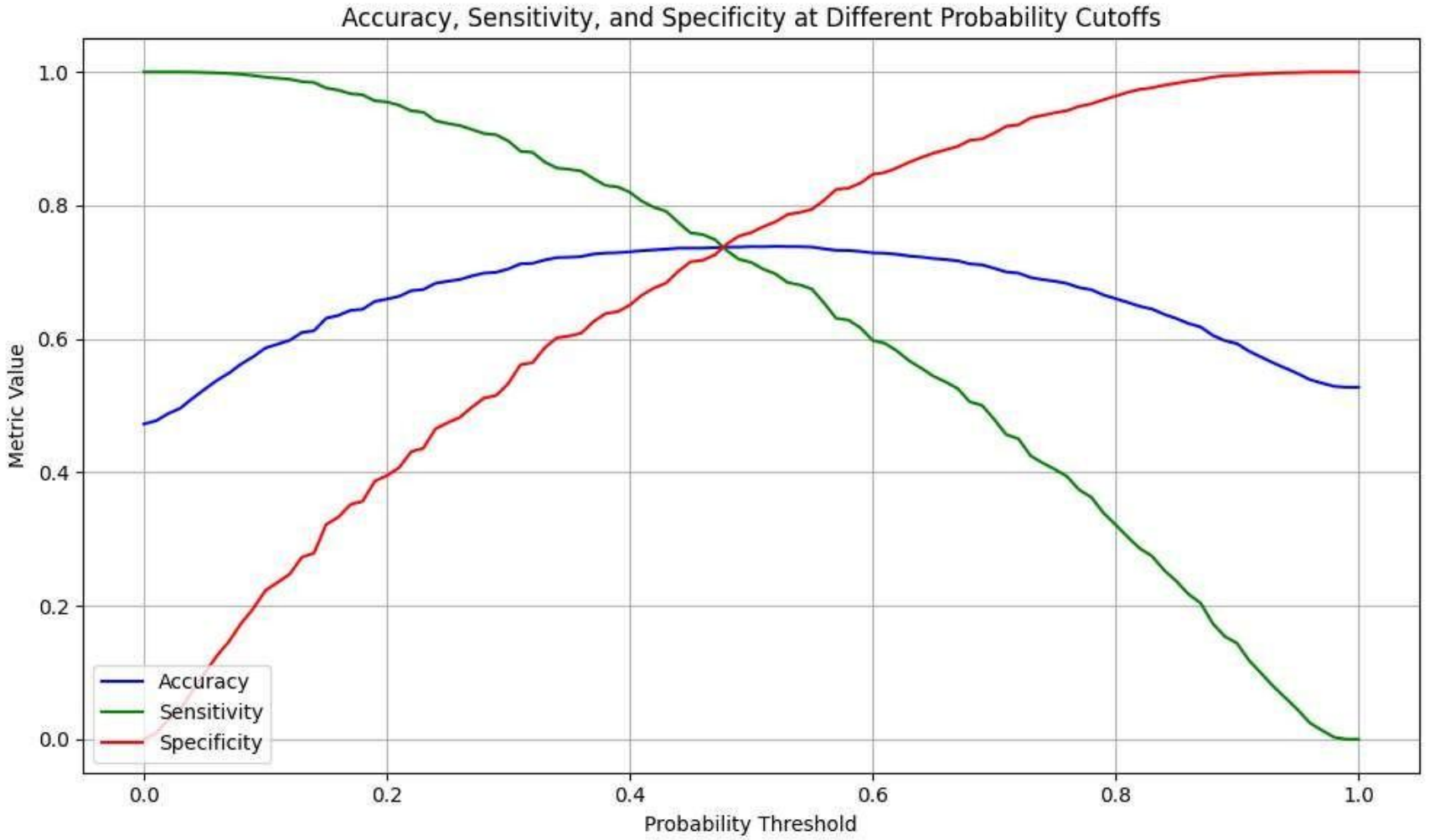
The confusion matrix serves as a foundational diagnostic tool, illustrating the model's performance on actual versus predicted outcomes. By dissecting true positives, false positives, true negatives, and false negatives, we gain critical insights into where the model excels and where it falters

ROC Curve – Significance



•**Evaluates Model Performance:** Shows the trade-off between true positive rate (sensitivity) and false positive rate.
•**Helps Choose Optimal Cutoff:** Aids in selecting the best probability threshold for classification.
•**AUC Value Interpretation:** The Area Under Curve (AUC) quantifies model's ability to distinguish between classes – closer to 1 means better performance.

Accuracy, Sensitivity, and Specificity at Different Probability Cutoffs



- **Model Accuracy:** Achieving a model accuracy of 74.03% indicates that a significant majority of predictions made by the model align with actual outcomes. This accuracy level reflects the model's overall effectiveness in distinguishing between employees who are likely to remain and those who may leave the organization.
- **Precision:** With a precision rate of 80%, the model demonstrates a high degree of reliability in its positive predictions. This means that when the model predicts an employee will leave, it is correct 80% of the time, suggesting a strong correlation between predicted turnover and actual turnover.
- **Recall:** A recall rate of 30.77% reveals a critical area for improvement, as it indicates that the model is only identifying a fraction of actual separations. This low sensitivity to true positives suggests a need to refine the model to better pinpoint employees at risk of leaving.
- **Specificity:** The model's specificity of 75.58% highlights its capacity to accurately identify employees who are likely to stay. This means that, while a majority of staying employees are correctly predicted, there is still room to enhance the precision of retention forecasts.
- **Sensitivity:** With a sensitivity measurement of 72.35%, the model's ability to detect true positives is relatively strong, yet it reflects a missed opportunity to identify and engage with more at-risk employees. Addressing this will be imperative for comprehensive retention strategies.

Improvement plan

- **Recommendations for Improvement:** In order to enhance retention outcomes, it is essential to incorporate qualitative data from employee feedback and further refine model parameters. Increasing the model's sensitivity and recall will enable more proactive interventions for employees identified as high-risk.