PREDICTING EMPLOYEE RETENTION
-Case Study

Using Logistic Regression

BY-

Khushboo Kumari

Kavya Makhija

Rajendran Karpaga Lakshmi

OBJECTIVE

- In this assignment we have be built a logistic regression model to predict the likelihood of employee retention based on the data such as demographic details, job satisfaction scores, performance metrics, and tenure.
- The aim of this 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 STEPS

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

DATA DICTIONARY

The data has 24 Columns and 74610 Rows. Following data dictionary provides the description for each column present in dataset:

- **Employee ID**: A unique identifier assigned to each employee.
- **Age**: The age of the employee, ranging from 18 to 60 years.
- **Gender**: The gender of the employee.
- **Years at Company**: The number of years the employee has been working at the company.
- **Monthly Income**: The monthly salary of the employee, in dollars.
- **Job Role**: The department or role the employee works in (e.g., Finance, Healthcare, Technology, Education, Media).
- **Work-Life Balance**: Perceived balance between work and personal life (Poor, Below Average, Good, Excellent).
- **Job Satisfaction**: Satisfaction with the job (Very Low, Low, Medium, High).
- **Performance Rating**: The employee's performance rating (Low, Below Average, Average, High).
- **Number of Promotions**: Total number of promotions received.
- **Overtime**: Number of overtime hours.
- **Distance from Home**: Distance between the employee's home and workplace (in miles).
- **Education Level**: Highest education level (High School, Associate Degree, Bachelor's, Master's, PhD).
- **Marital Status**: Marital status (Divorced, Married, Single).
- **Number of Dependents**: Number of dependents the employee has.
- **Job Level**: The job level (Entry, Mid, Senior).
- **Company Size**: Size of the company (Small, Medium, Large).
- **Company Tenure (In Months)**: Total number of months the employee has worked in the industry.
- **Remote Work**: Whether the employee works remotely (Yes or No).
- **Leadership Opportunities**: Availability of leadership opportunities (Yes or No).
- **Innovation Opportunities**: Availability of innovation opportunities (Yes or No).
- **Company Reputation**: Employee's perception of company reputation (Very Poor to Excellent).
- **Employee Recognition**: Level of recognition received (Very Low, Low, Medium, High).
- **Attrition**: Whether the employee has left the company.

DATA UNDERTSANDING & CLEANING

- The data has 24 columns and 74610 rows.
- While handling missing values, we found that the columns 'Distance from Home' and 'Company Tenure(In Months)' have 2.56% and 3.23% missing values respectively. Since the percentage of missing is very small, we removed the rows.
- The duplicate data was identified and dropped from the dataset.

HANDLE REDUNDANT VALUES

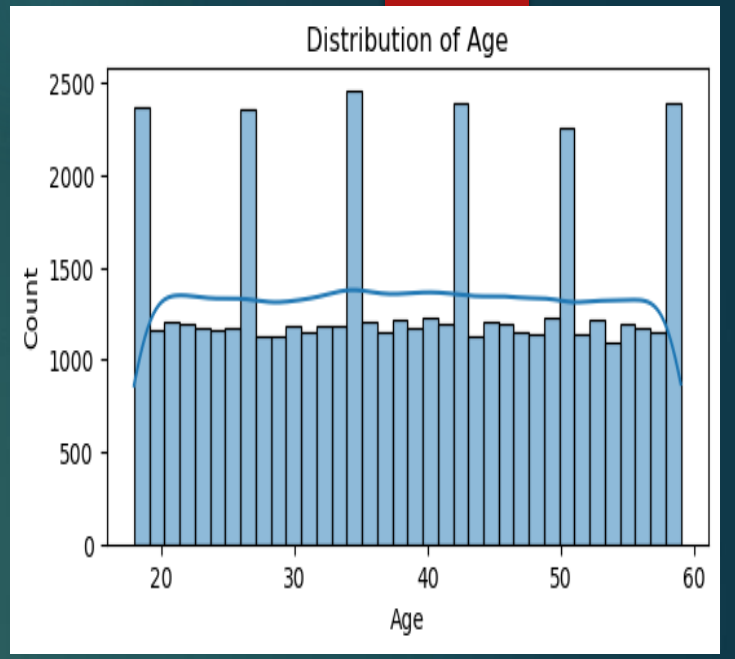
- The following columns can be removed
- Gender can be removed as it is a protected attribute also it should be removed to avoid bias
- Marital Status: Another protected attribute with limited value
- Leadership Opportunities: Majority (Almost all) "No". This is not very useful
- Innovation Opportunities: Majority (Almost all) "No". This is not very useful
- Employee ID: It's just a unique identifier. it doesn't hold any meaningful relationship to attrition.

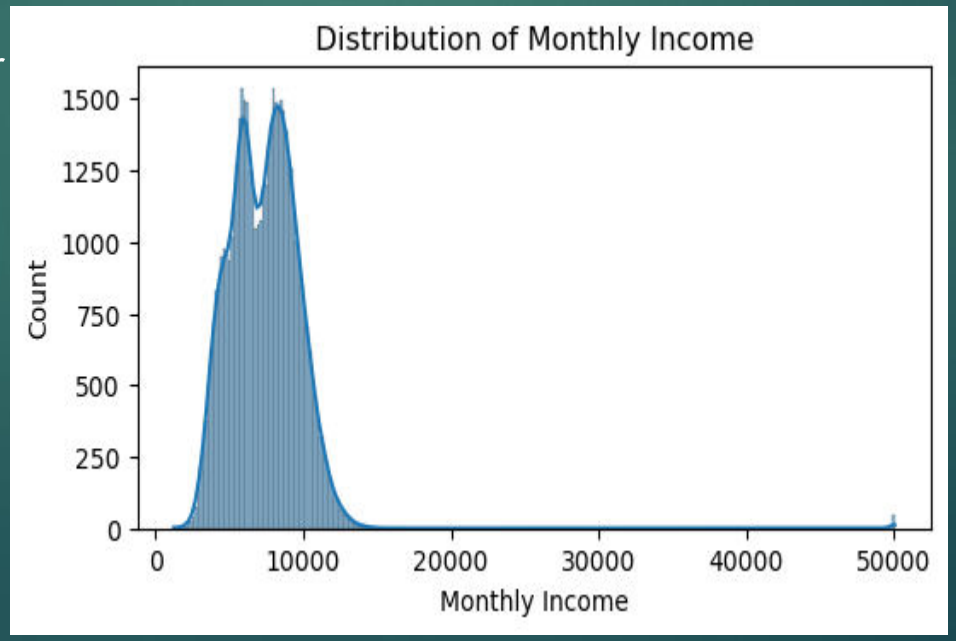
TRAIN-VALIDATION SPLIT

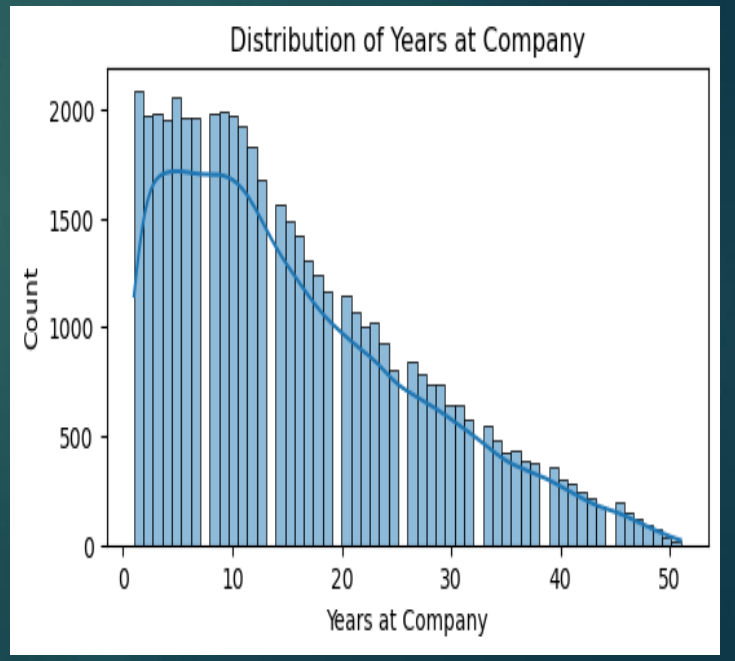
To prepare the data for model building, all feature variables are stored in variable X by dropping the columns ‘Attrition’. The target variable y was assigned the ‘Attrition’ column. The dataset was split into training and validation set using a 70-30 split, where 70% of data was used for training and rest 30% for validation.

EDA ON TRAINING DATA: UNIVARIATE ANALYSIS

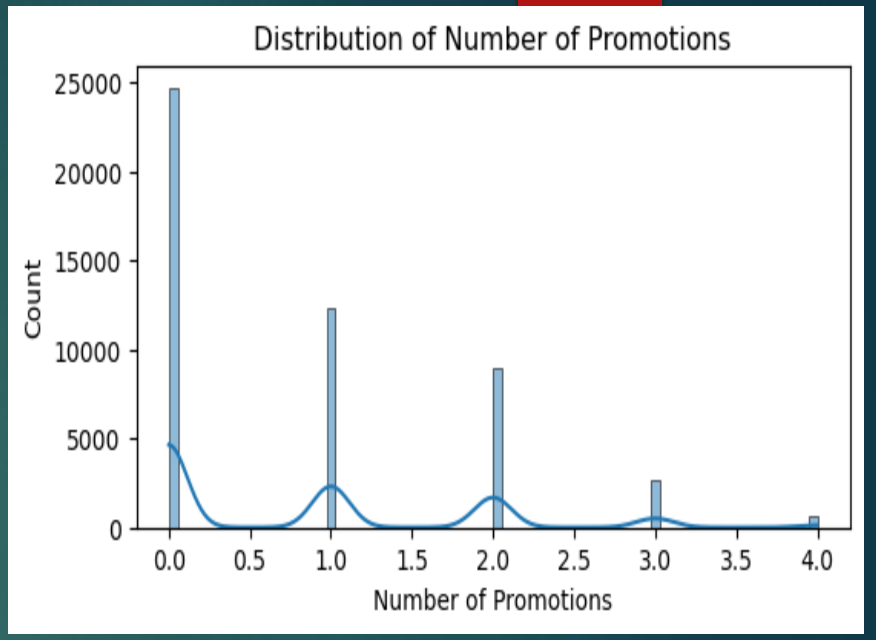
- In the histplot of column 'Age' it is visible that the distribution is uniform with exception of the peaks, suggesting a generally even spread of ages. There are notable peaks at ages around 20,30,40,50 and 60.
- In the column 'Years at Company' we see a decreasing trend indicating that the number of employees decreased as tenure increases.
- The 'Monthly Income' histplot illustrates a population where the majority earns relatively low monthly income, with few individuals earning substantially more.

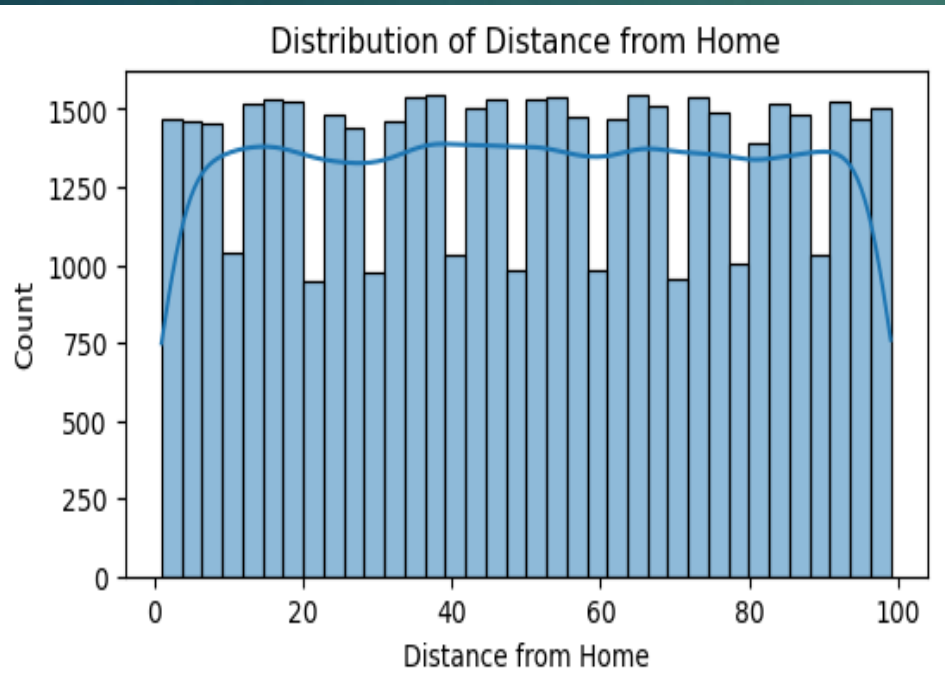


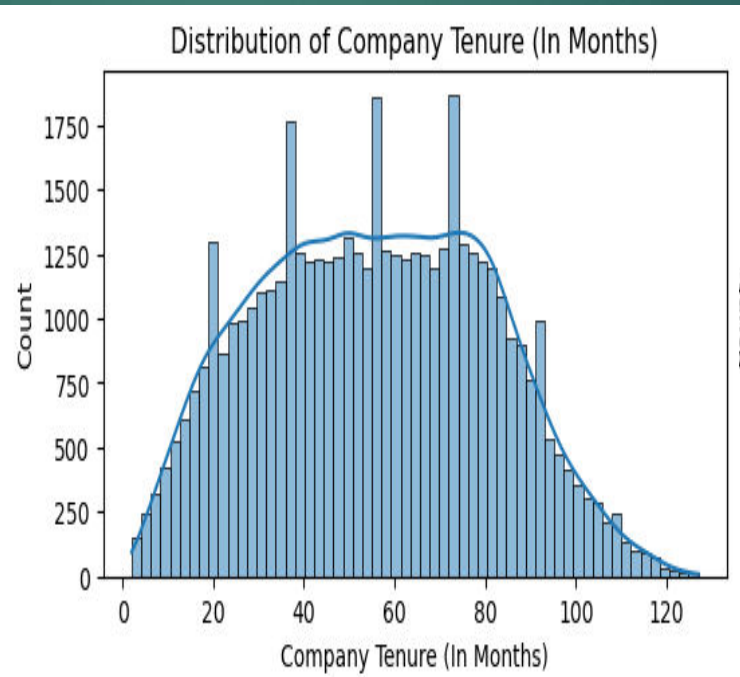


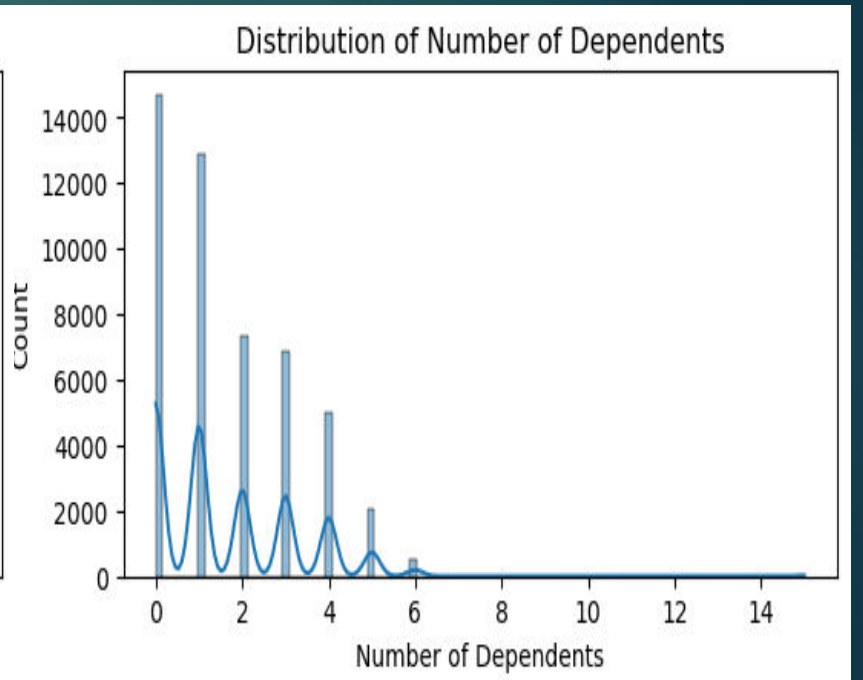


The histplot of ‘Number of promotions’ suggests that the employees who receive more promotions are less likely to leave the company.
In the histplot of ‘Number of Dependents’ it appears that employees with fewer dependents are more likely to leave the company.
‘Distance from Home’ graph indicated a spike at the 50-unit distance. This suggests housing, transport and work-life balance issue hence increasing turnover.

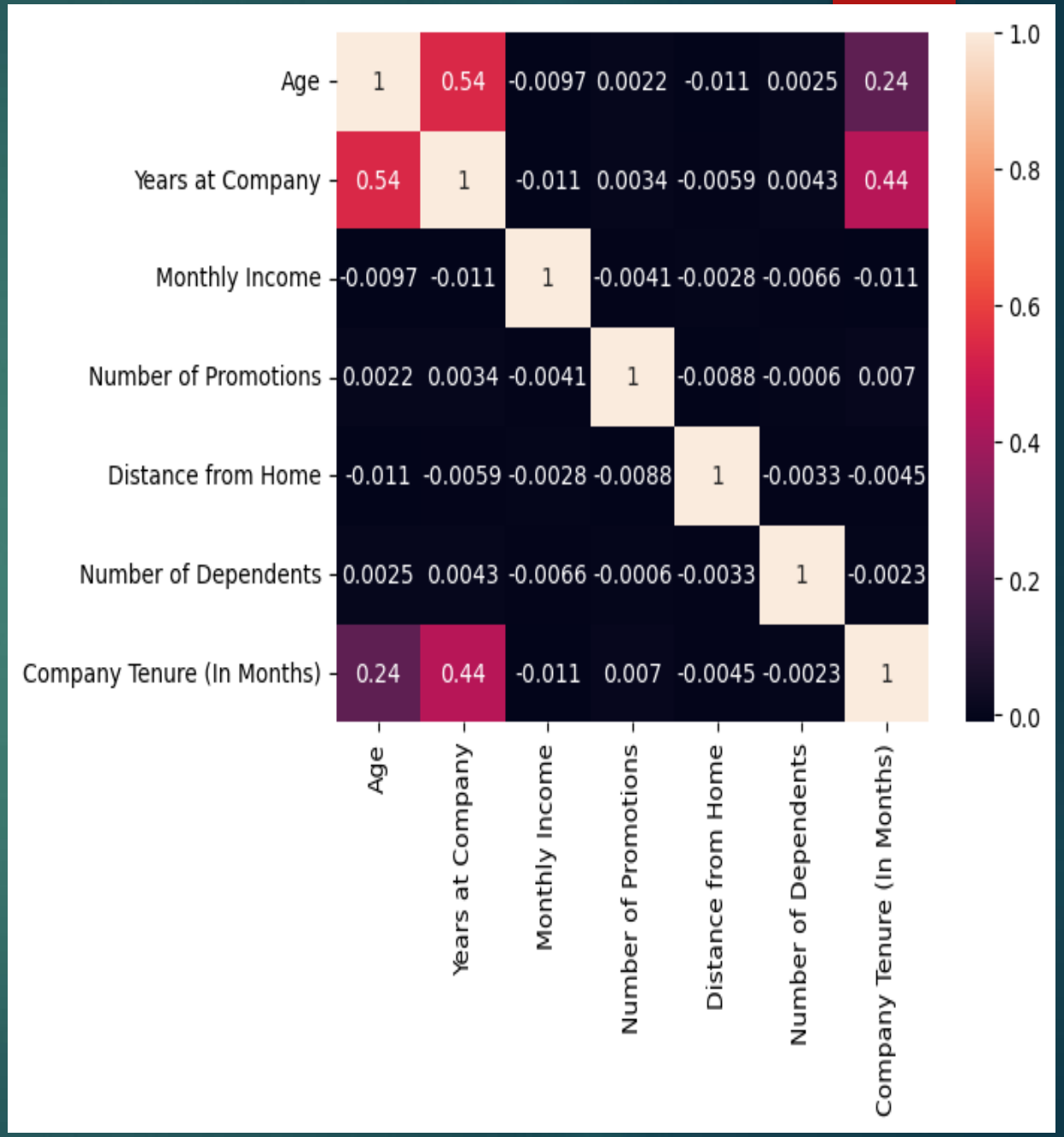






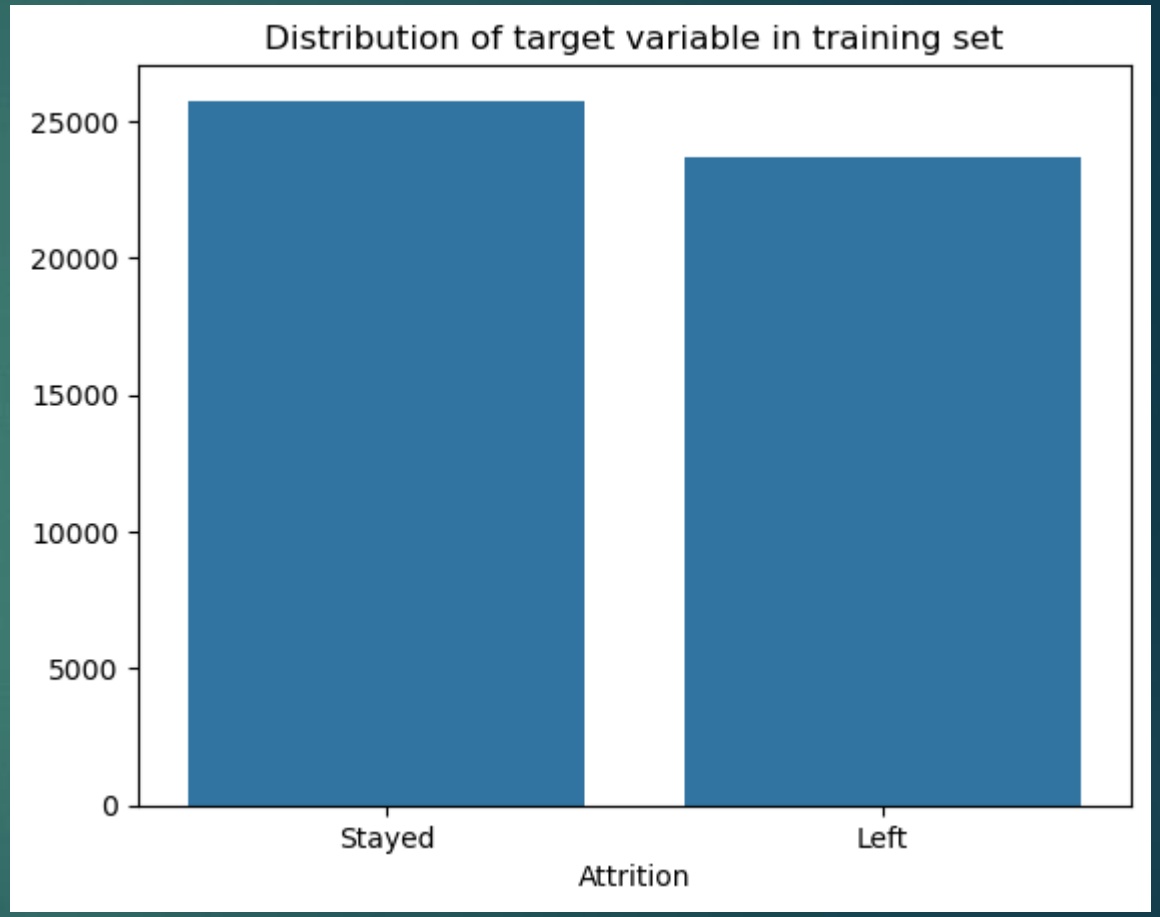


- Correlation coefficients range from -1 to 1, whereas values close to 1 or -1 indicate a strong correlation and values close to 0 indicate a weak correlation.
- The graph depicts the correlation matrix of numerical columns.
- From the graph, it is depicted that the columns ‘Age’ and ‘Years at Company’ are highly correlated.



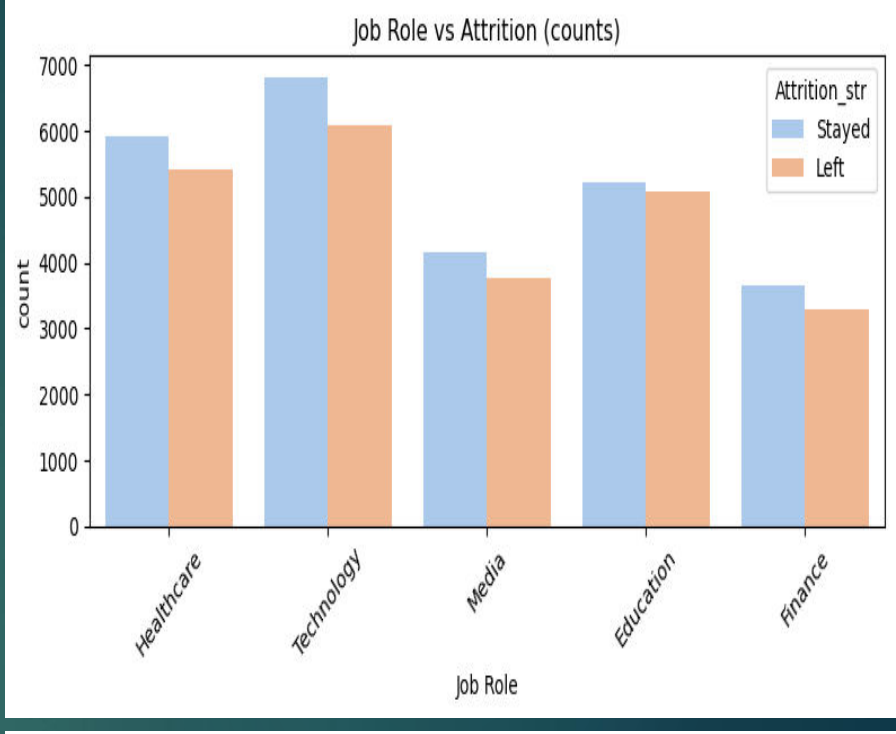
CLASS BALANCE

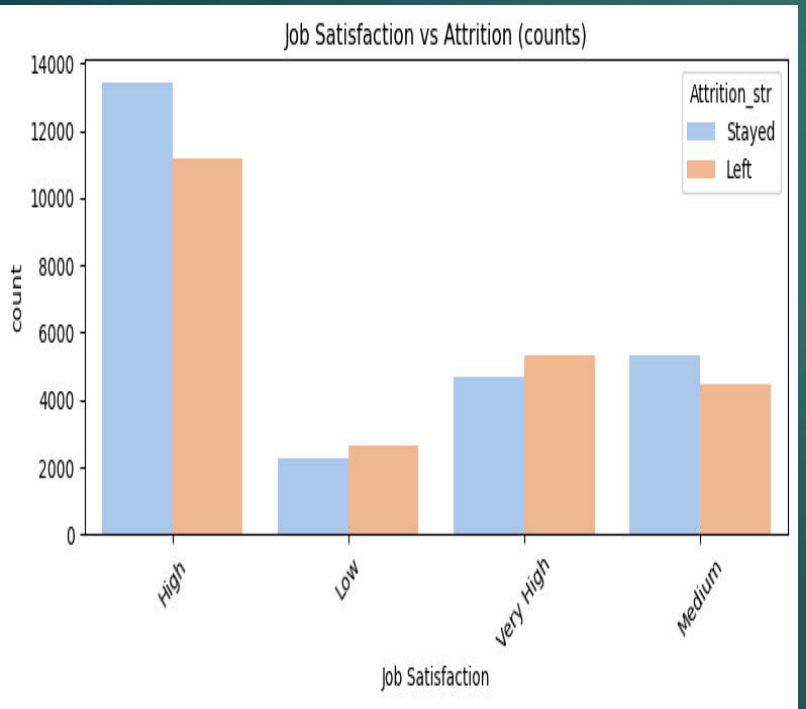
- The graph depicts the class balance of the target variable in training set.
- The ‘Stayed’ count is 25745 and ‘Left’ count is 23645, indicating that the classes are balanced.

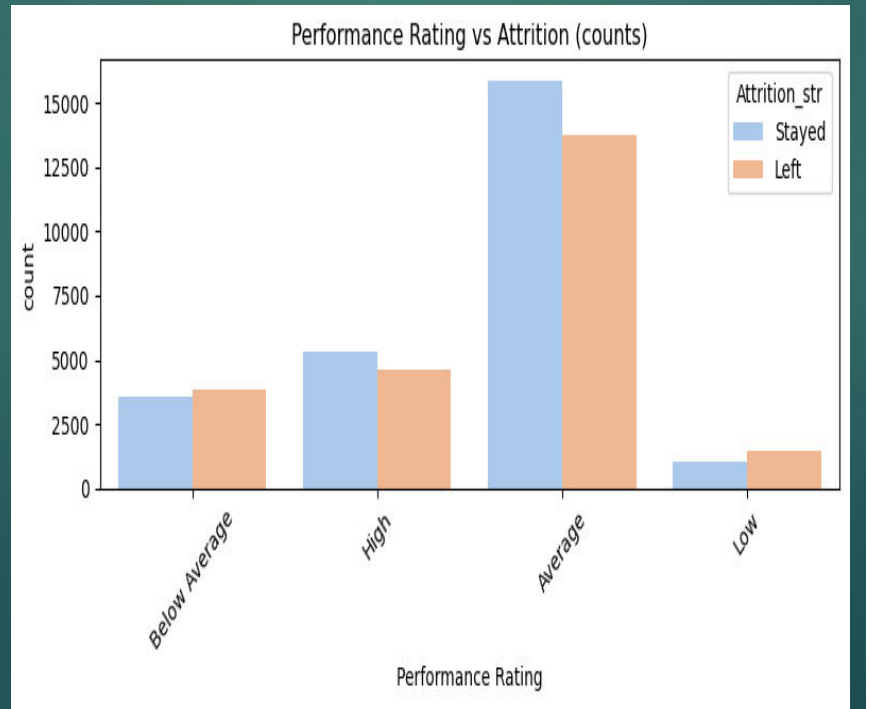


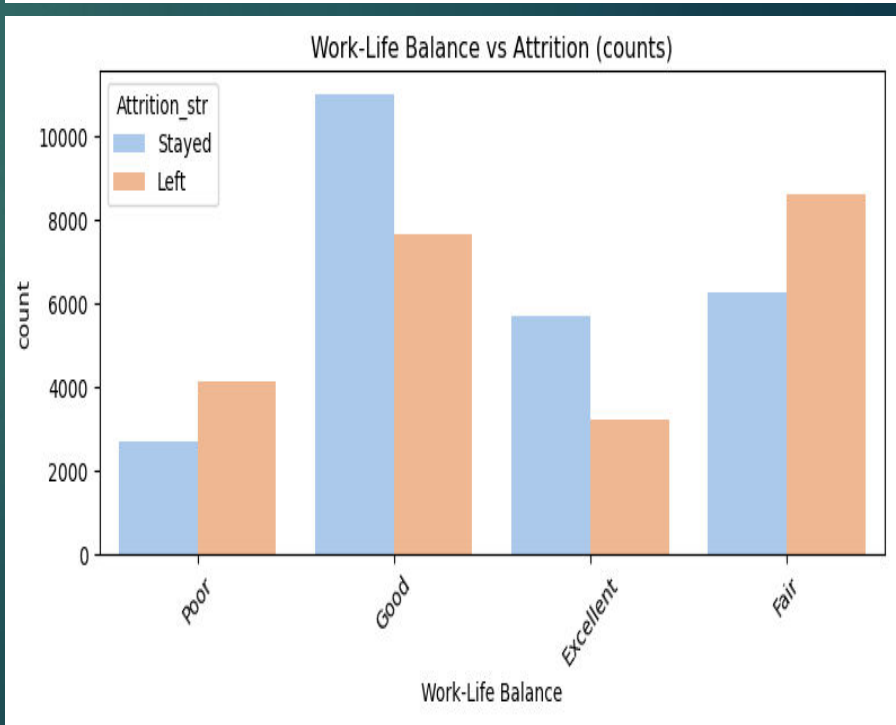
BIVARIATE ANALYSIS

- The graph between ‘Job Role’ and ‘Attrition’ depicts ‘Technology’ has highest number of employees leaving with a significant gap between those who stayed and those who left.The other sectors experience more stable workforce retention rates.
- ‘Work-Life Balance’ vs ‘Attrition’ graph illustrates the highest number of employees who stayed reported having a good work-life balance, with significantly lower number leaving. The graph suggest a strong correlation between work-life balance and employee retention
- Graph between ‘Performance Rating’ and ‘Attrition’ depicts ‘Average’ has highest number of employees leaving with a significant gap between those who stayed and those who left.
- ‘Job Satisfaction’ vs ‘Attrition’ graph illustrates the highest number of employees who stayed reported having a high job satisfaction.

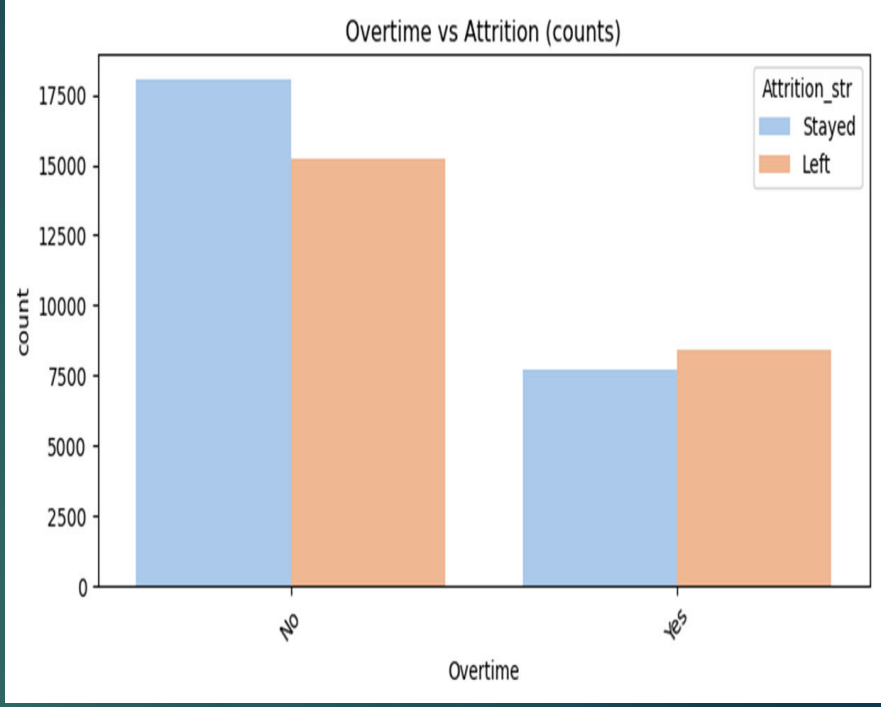


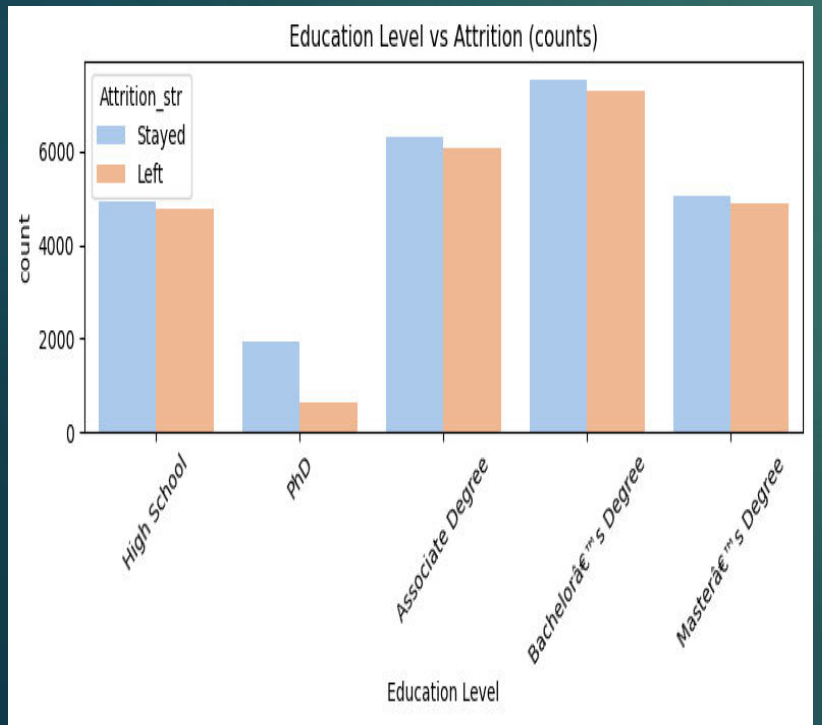


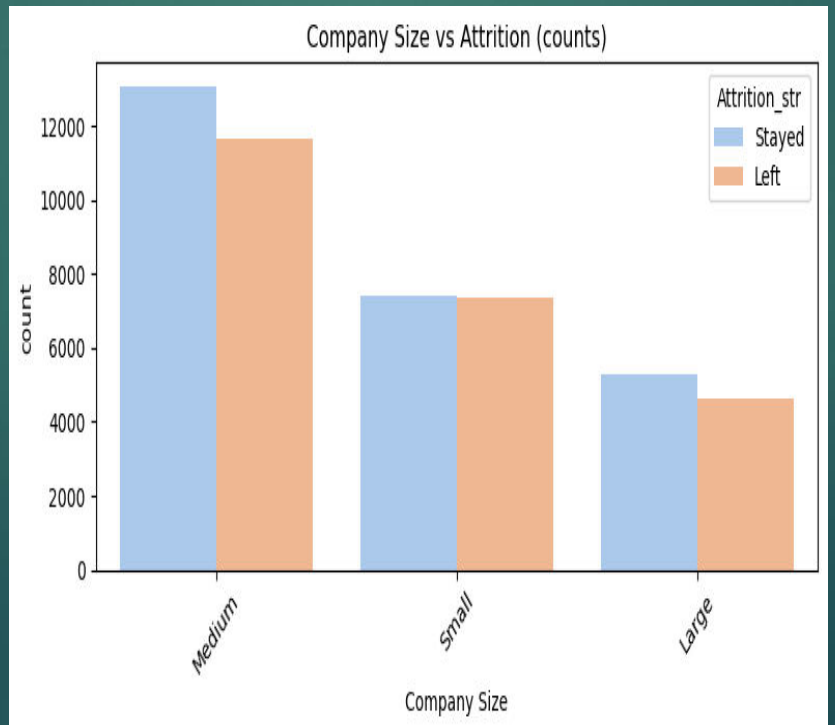


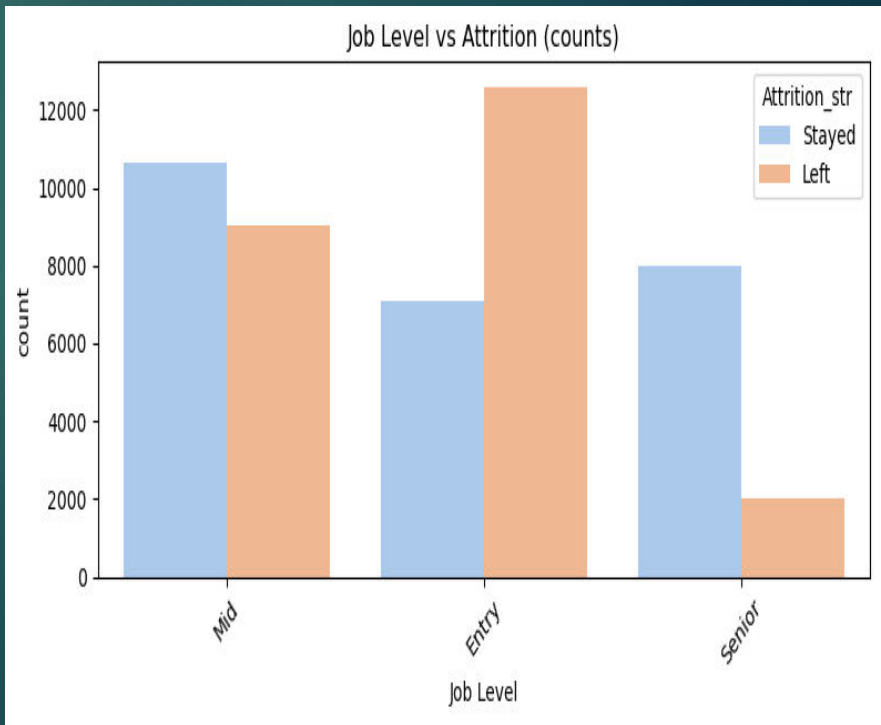


- The graph between ‘Overtime’ and ‘Attrition’ suggests that overtime work maybe a contributing factor to employee attrition.
- The graph ‘Job Level’ vs ‘Attrition’ suggests that attrition is highest at entry level, followed by mid-level and is lowest at senior level.
- The ‘Company Size’ vs ‘Attrition’ graph depicts that employee attrition is more prevalent in medium-sized company, while large companies have the lowest attrition rates.
- The graph between ‘Education’ and ‘Attrition’ suggests that attrition rates is higher among employees with high school education compared to those with higher degrees.

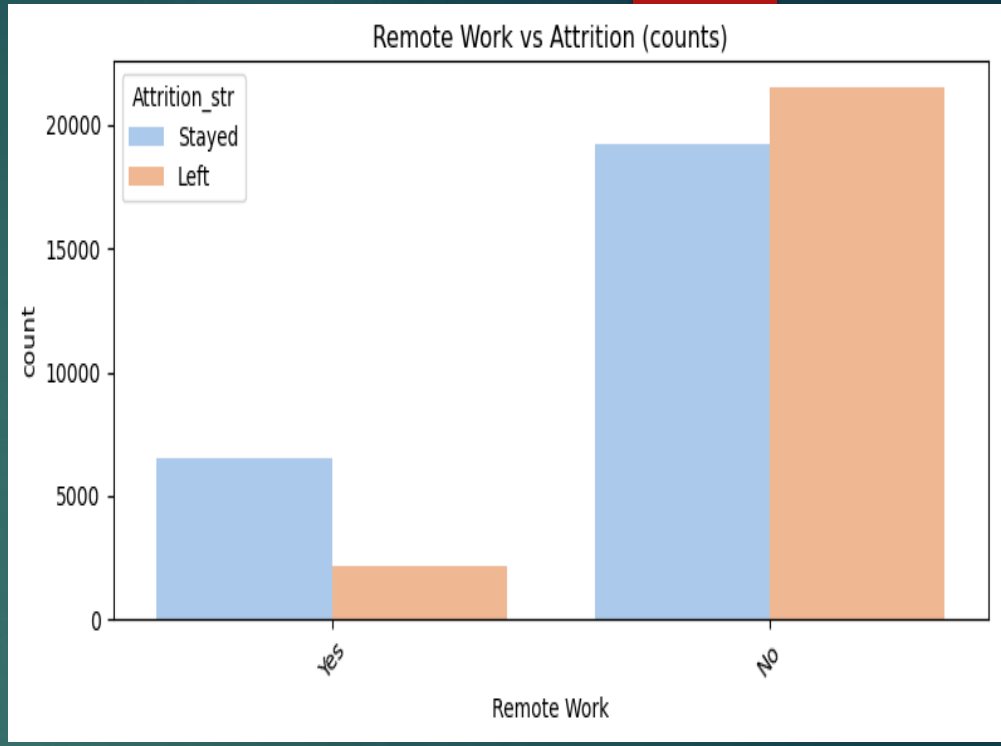


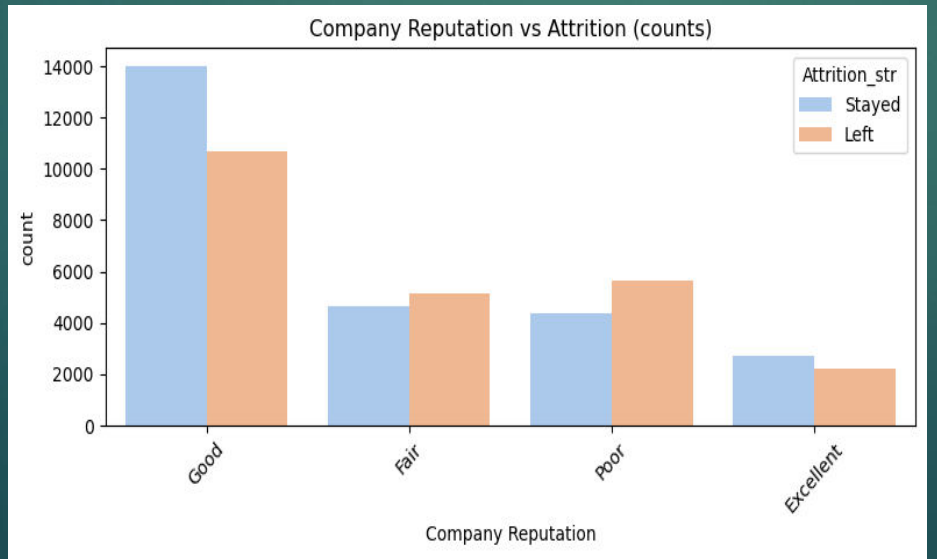


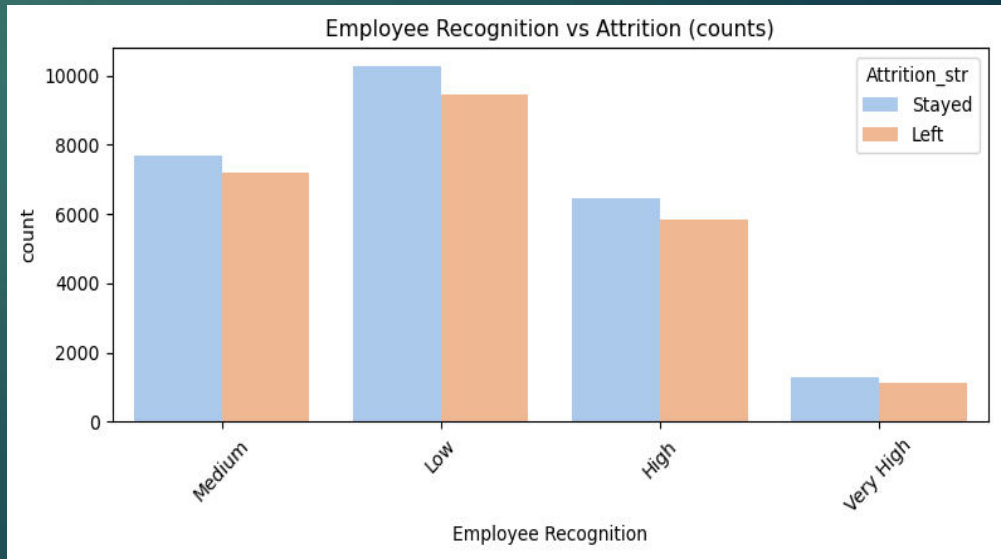




- The employees who work remotely are more likely to stay
- ‘Employee Recognition’ vs ‘Attrition’ graph suggests a clear inverse relationship between employee recognition and attrition. Low recognition is a strong predictor of employee turnover, while high recognition is a strong factor in employee retention.
- The graph between ‘Company Reputation’ and ‘Attrition’ suggests a positive correlation. Companies with good and excellent reputations experience significantly lower attrition rates compared to those with fair and poor reputations.







FEATURE ENGINEERING

- DUMMY VARIABLE CREATION:

The first step is to convert categorical columns to numeric using get_dummies with drop_first=True to avoid multicollinearity. Then the original categorical columns were removed from train and validation sets. The processed dummy and numeric columns were then combined into a dataset.

FEATURE SCALLING:

To ensure numeric features are on a consistent scale, StandardScaler was applied. The numerical columns in the training data were scaled using fit_transform(). The same method was applied to validation set. during training. This improves model performance by treating all features on the same scale.

MODEL BUILDING

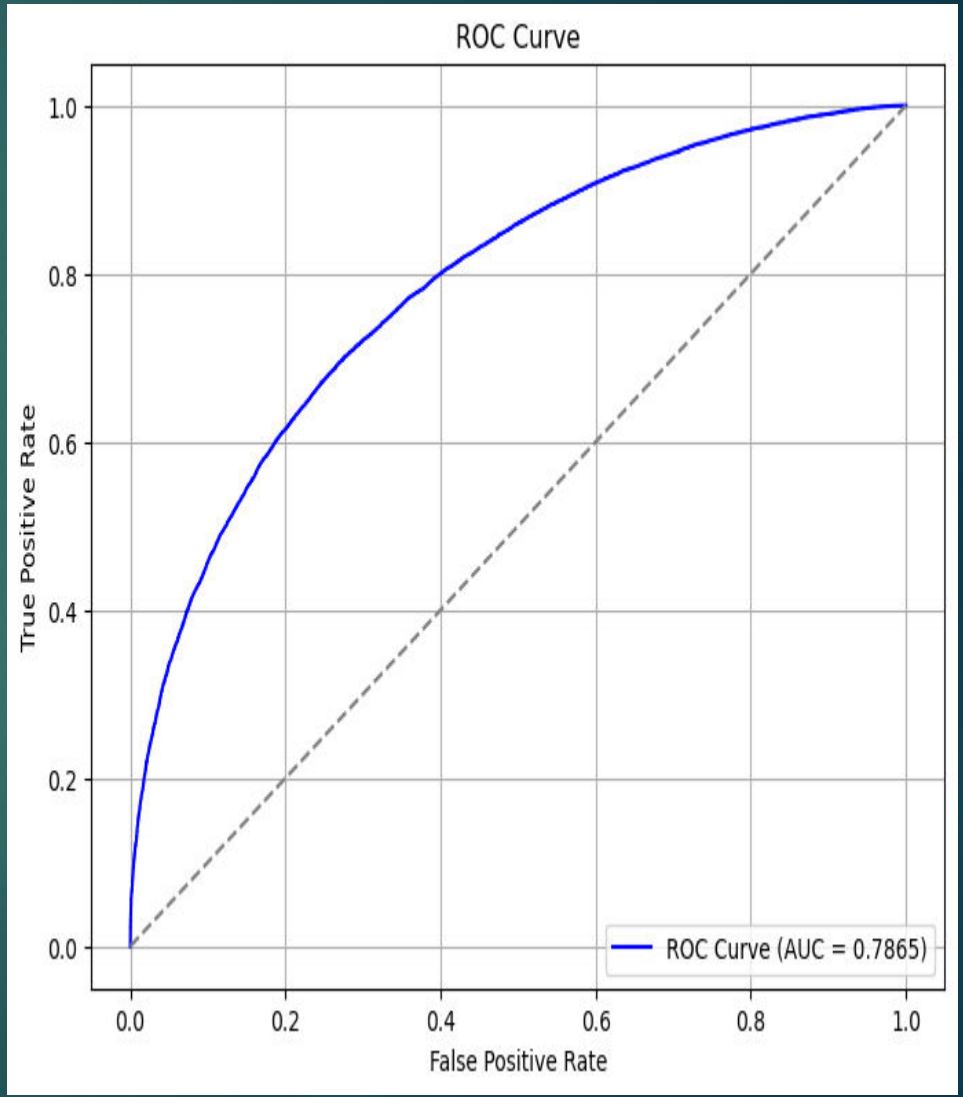
- Recursive Feature Elimination(RFE) was used to select top 15 important features for logistic regression model. The model was built using statsmodels. VIF was used to check multicollinearity. All VIFs are well below 5, most are even around 1.0, which is excellent. This logistic regression coefficients and p-values were trustworthy so there was no need to drop or transform variables due to multicollinearity. The overall accuracy of model was 0.7109. This means the model correctly predicted whether employees stayed or left about 71.09% of the time.
- To evaluate model's performance, a confusion matrix was created based on predictions from training set. The matrix contains four key values True Negatives(TN), False Positive(FP), False Negatives(FN), True Positives(TP). Using this we calculate :

  Sensitivity = 71.53%

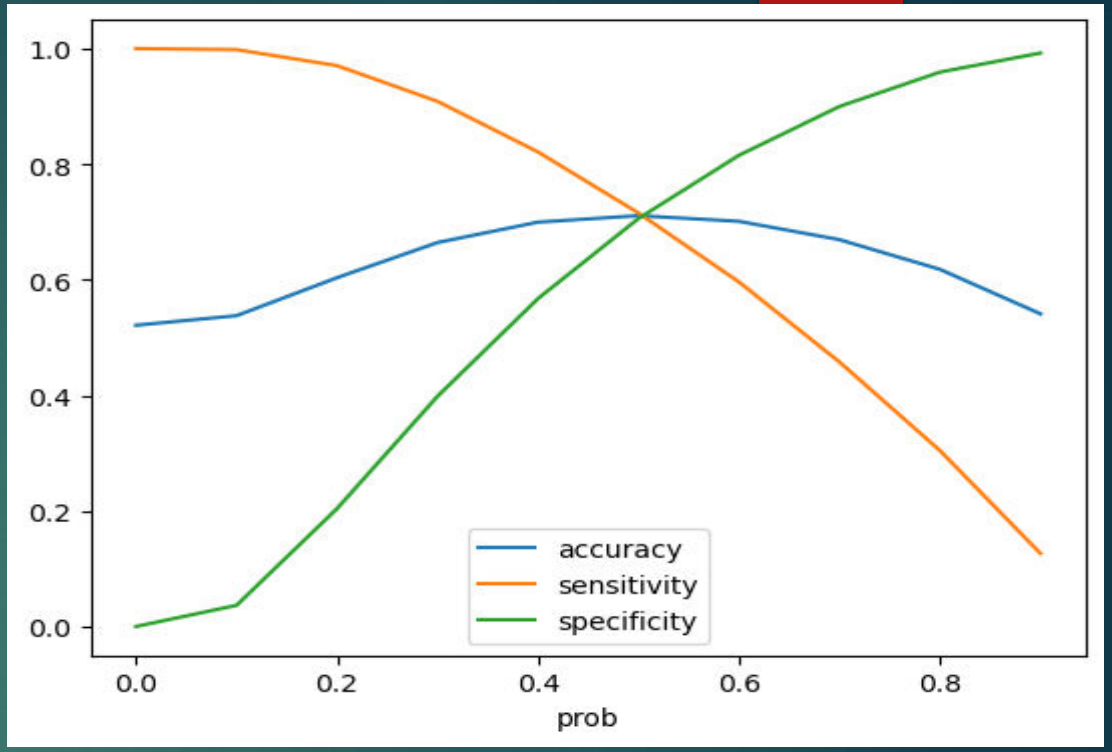
  Specificity = 70.55%

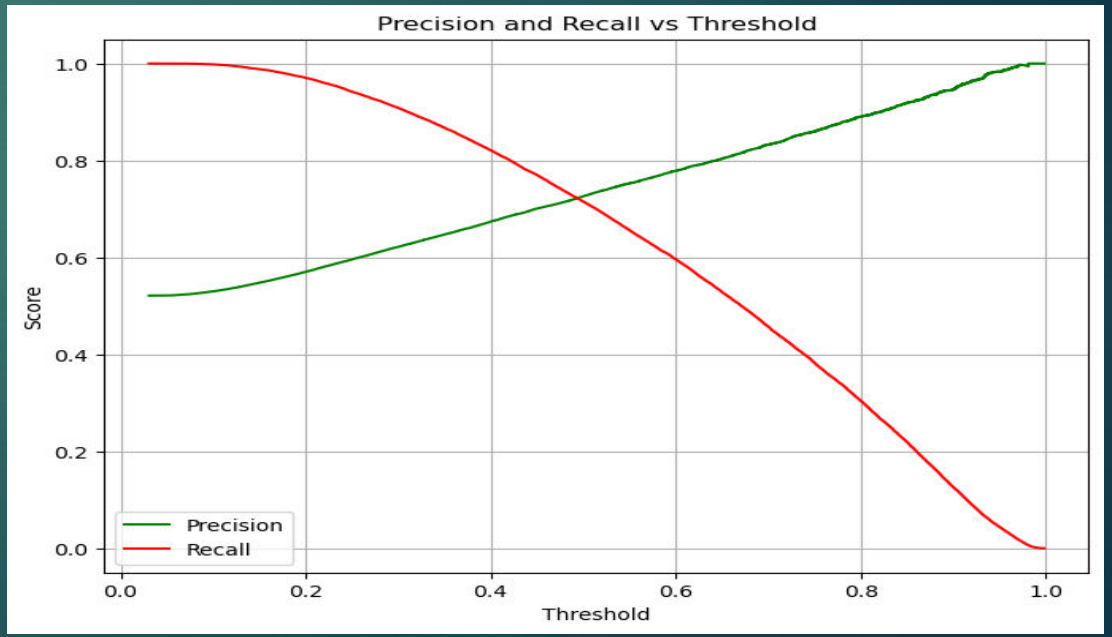
  Precision = 72.59%

  Recall = 71.53%
- To improve model's performance beyond the default 0.5 threshold, we analyses ROC curve. The AUC curve was 0.7865. An AUC of 0.78 means the model has a good ability to distinguish between classes — definitely better than random (which is 0.5), and not far off from the 0.8+ "strong" range.



- To determine the best probability threshold for classification, predictions were evaluated at different cutoff values ranging from 0.0 to 0.9. From this we calculated accuracy, sensitivity and specificity.
- From the graph,0.5 is the optimum point to take it as a cutoff probability.
- A Precision-Recall curve was plotted to evaluate model performance across different probability threshold.
- The model's precision and recall were plotted against various probability thresholds to visualize how both metrics change with the cutoff.
- The graph between 'Precision and Recall' and 'Threshold' indicates that adjusting the threshold impacts the performance of classification model. Lower threshold favors recall, while a higher favors precision.





PREDICTION AND MODEL EVALUATION

- Logistic regression model was trained using scaled key features. The predictions were made on validation set using optimal cutoff ,i.e., 0.5. A confusion matrix was created to better understand model's classification performance. The overall accuracy is 70.97%. Using confusion matrix we calculated,

  Sensitivity (Recall) = 71.17%: The model correctly identifies 71.17% of employees who actually stay.

  Specificity = 70.74%: It accurately identifies 70.74% of employees who leave.

  Precision = 72.56%: Of those predicted to stay, 72.56% actually stay, indicating good prediction accuracy for stayers.

  Recall = 71.17%: Similar to sensitivity, showing moderate success in identifying stayers, though some are missed (False Negatives).

  Overall, the model does a moderate job at predicting employee retention

CONCLUSION/KEY INSIGHTS

- Employees with fewer promotions are more likely to leave. Hence, the company should implement mentorship programs, transparent career paths etc.
- Attrition is high among employees who earn low monthly income. The company should update salary packages to make sure they are fair.
- The company should provide early career development as the highest attrition is seen at entry level and mid-level positions.
- Job satisfaction and performance rating have strong relation with attrition. So, regular surveys, manager feedback can boost satisfaction and performance.
- Employees working overtime are more likely to leave. The company should promote work-life balance.
- Providing or expanding remote work options could be a strategy to improve retention.
- Medium-sized companies have higher attrition.