Employee Retention Case Study

By: Pratham Kumar, Pranjil Martin and Prakhar Gupta

PROBLEM STATEMENT

Business Objective

A mid-sized technology company wants to improve its understanding of employee retention to foster a loyal and committed workforce. While the organization has traditionally focused on addressing turnover, it recognises the value of proactively identifying employees likely to stay and understanding the factors contributing to their loyalty.

- In this assignment we'll be building 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 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.

METHODOLOGY AND TECHNIQUES

1. Data Understanding

   Load the data and understand the basic statistical summary of the data

2. Data Cleaning

   2.1 Handle the missing values

   2.2 Identify and handle redundant values within categorical columns

   2.3 Drop redundant columns

3. Train-Validation Split

   3.1 Split the data into train and validation with a 70:30 ratio

4. EDA on Training Data

   4.1 Perform univariate analysis

   4.2 Perform correlation analysis

   4.3 Check the class balance

   4.4 Perform bivariate analysis

5. Feature Engineering

   5.1 Dummy variable creation

   5.2 Feature scaling

6. Model Building

   6.1 Feature selection

   6.2 Building a logistic regression model

   6.3 Find the optimal cutoff

7. Prediction and Model Evaluation

   7.1 Make predictions over the validation set

   7.2 Calculate the accuracy of the model

   7.3 Create a confusion matrix and create variables for true positive, true negative, false positive and false negative

   7.4 Calculate sensitivity and specificity

   7.5 Calculate precision and recall

VISUALIZATION

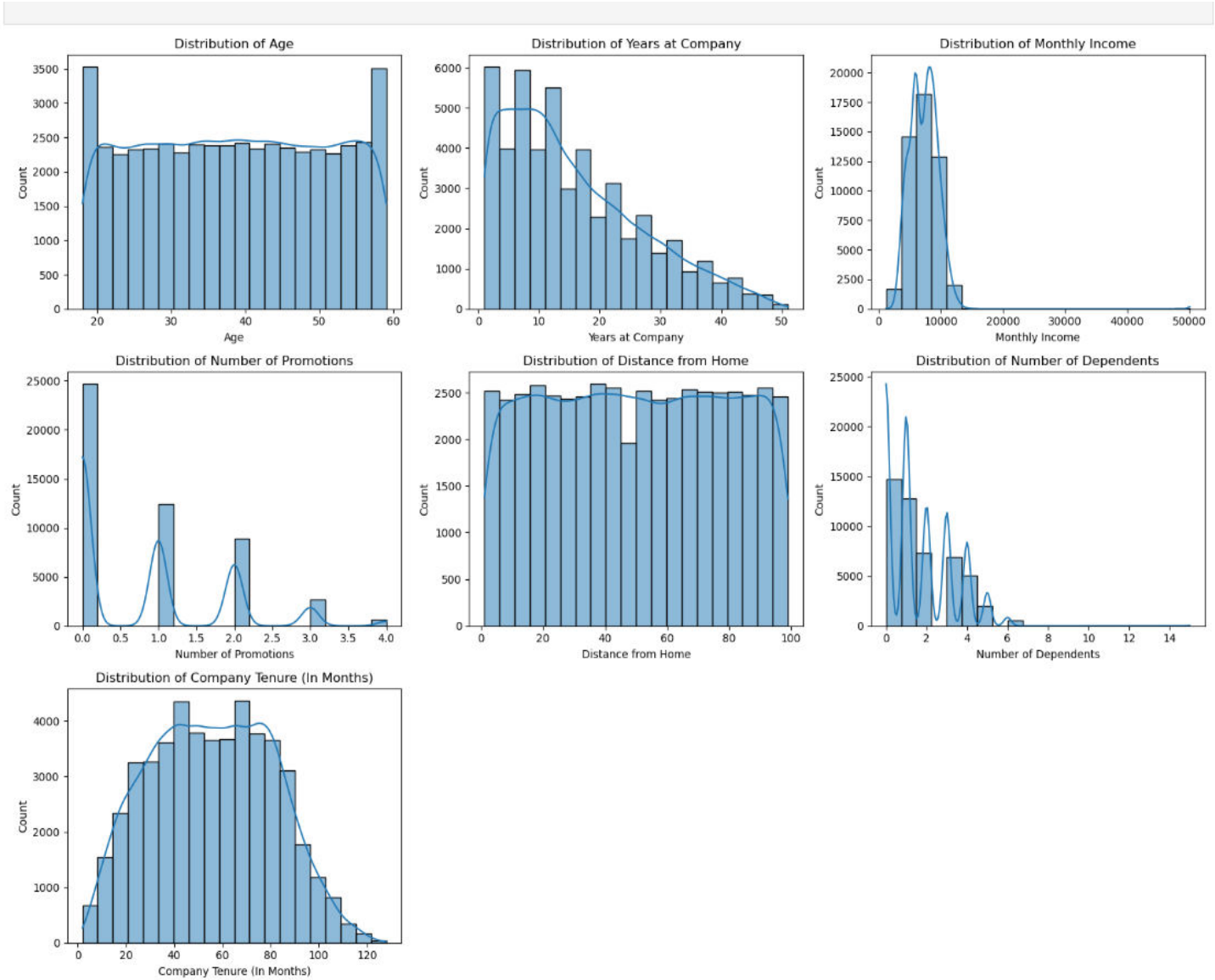
- Univariate Analysis(Numerical)

1.Age: employee age ranging (18-60 years old).

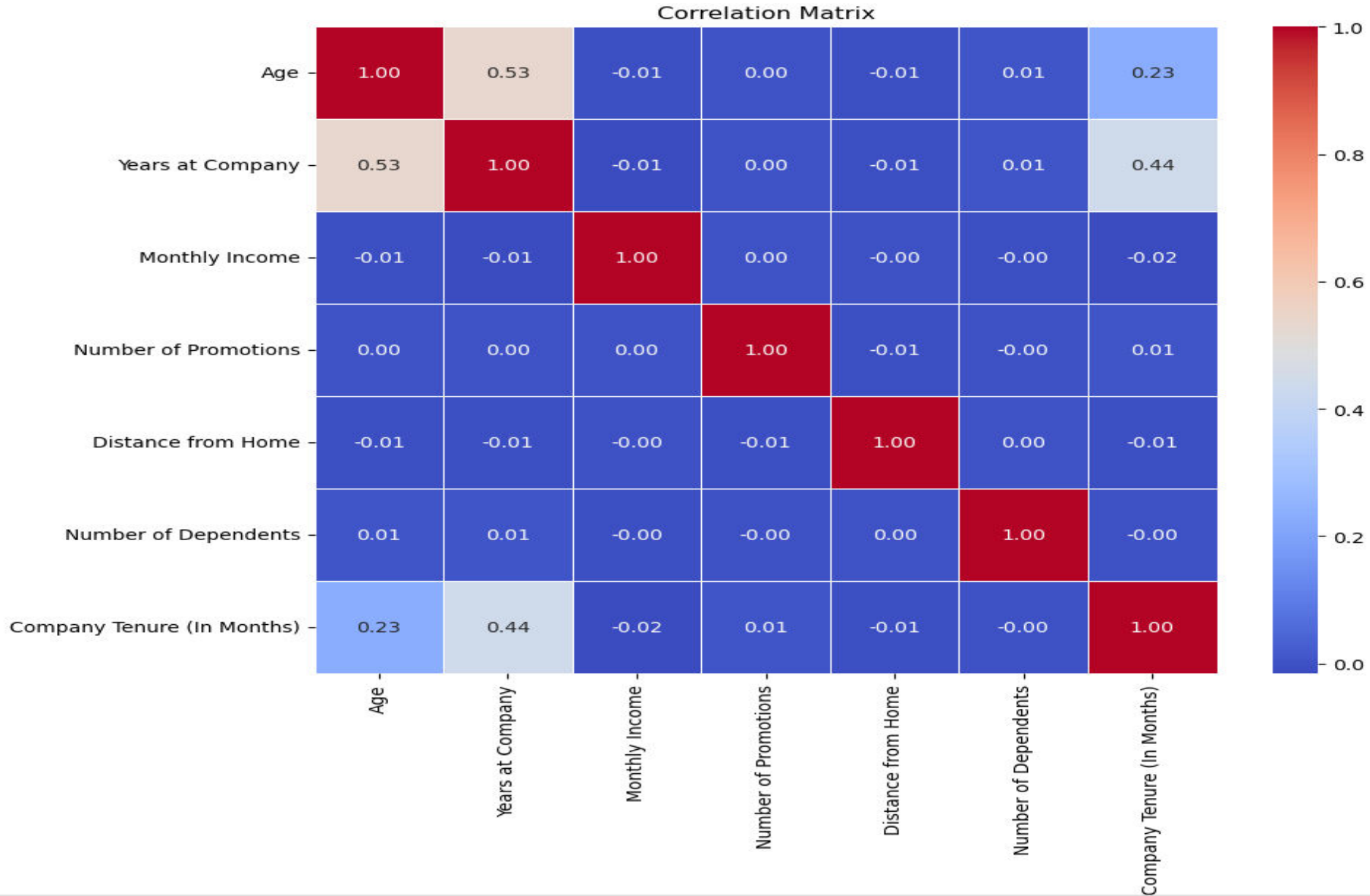
2.Most Number of employees have monthly income below 10,000 dollars.

3.Most number of employees have 0 promotions.

4. Most of employees have 0-2 dependents.



CORRELATION



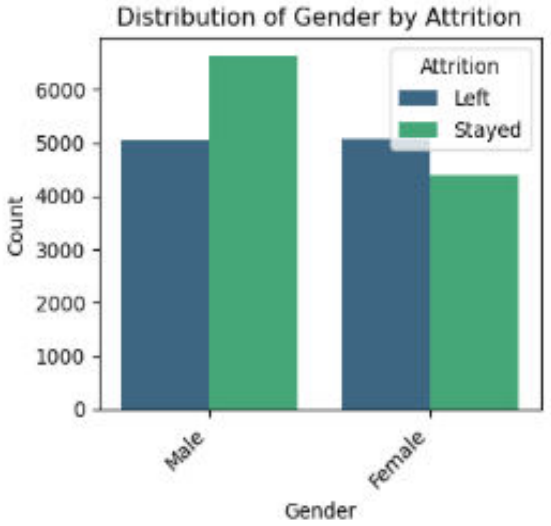
CLASS BALANCE

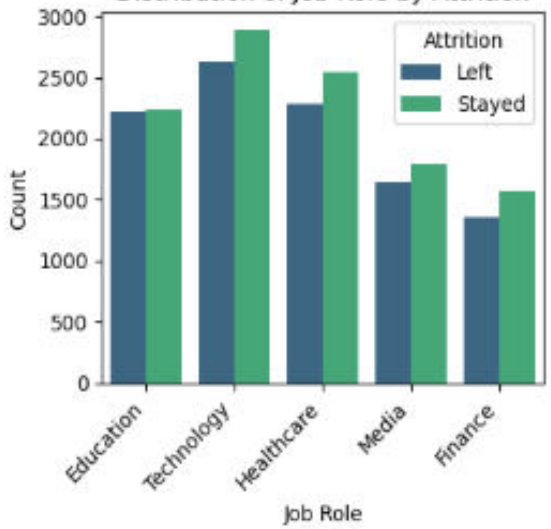
```
Class Distribution:
Attrition
Stayed    11032
Left      10136
Name: count, dtype: int64

Class Distribution (Normalized):
Attrition
Stayed    0.521164
Left      0.478836
Name: proportion, dtype: float64
```

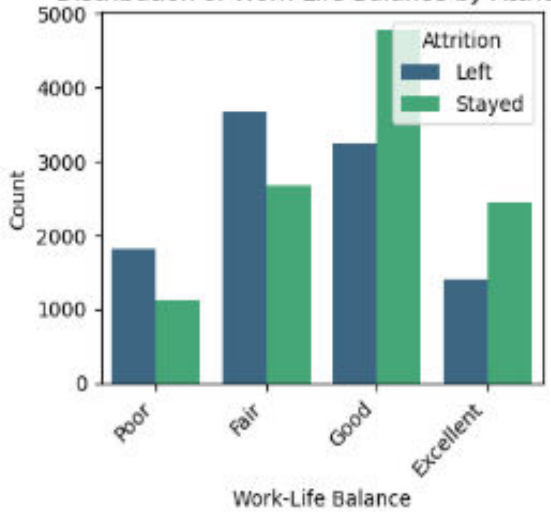
Attrition Rate & Class Distribution

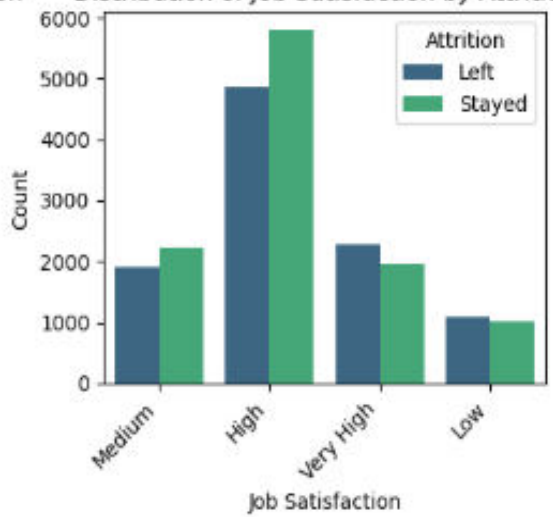
- **52.12% of employees stayed**, while **47.88% left**.
- The dataset is **reasonably balanced**, meaning models won't be heavily biased towards one class.

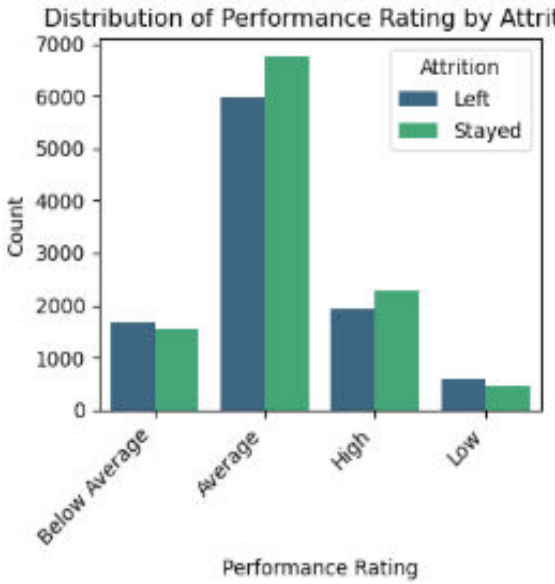
BIVARIATE ANALYSIS(CATEGORICAL)

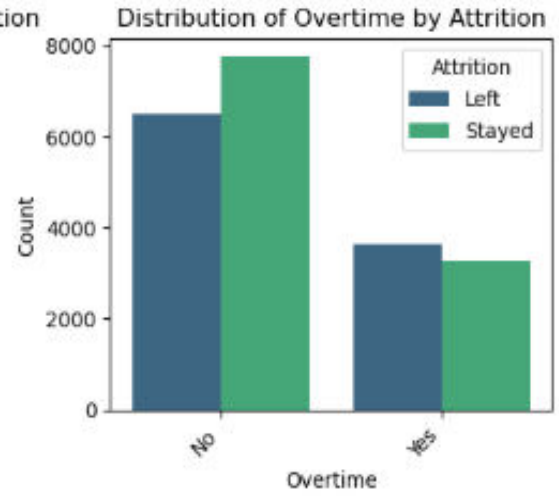


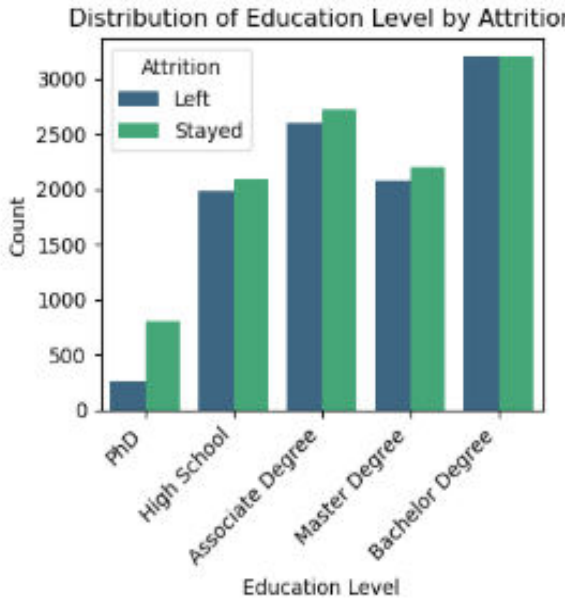


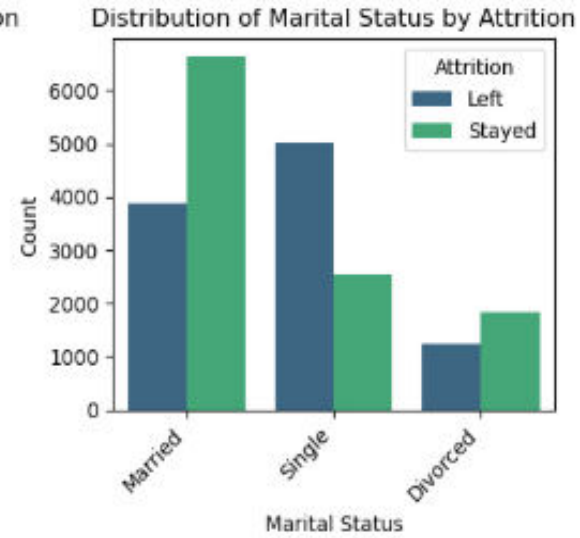




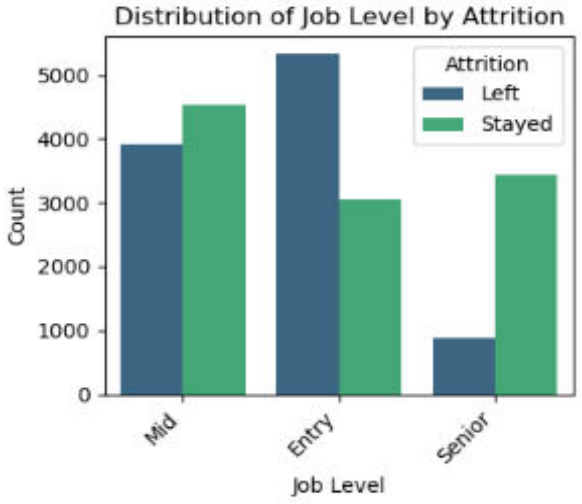


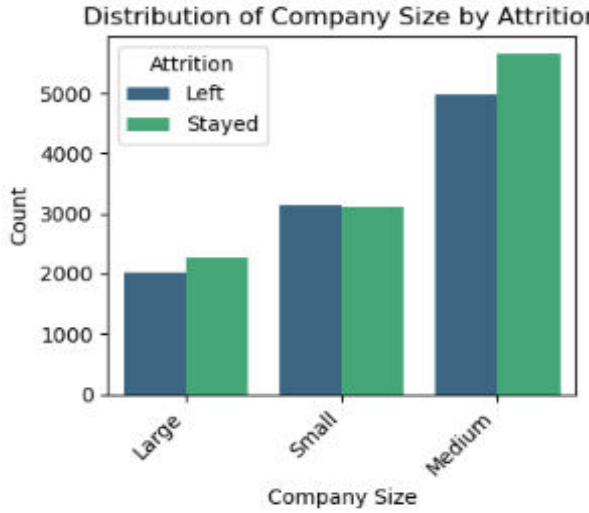


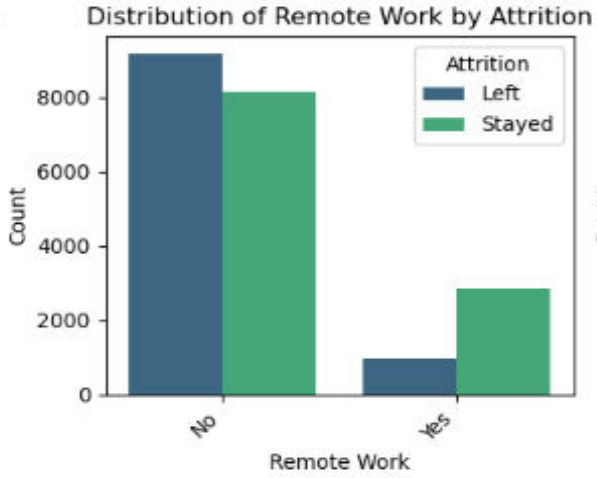


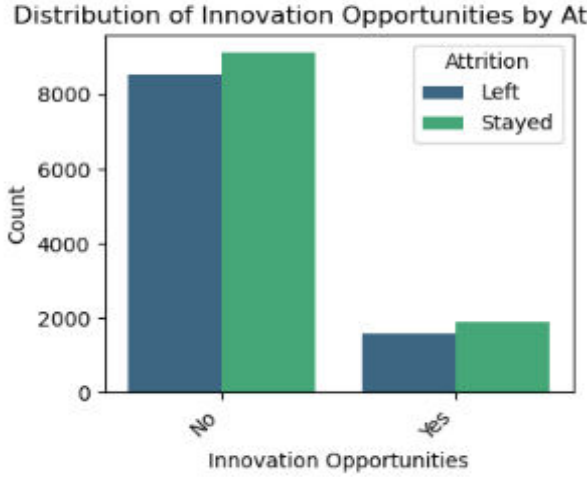


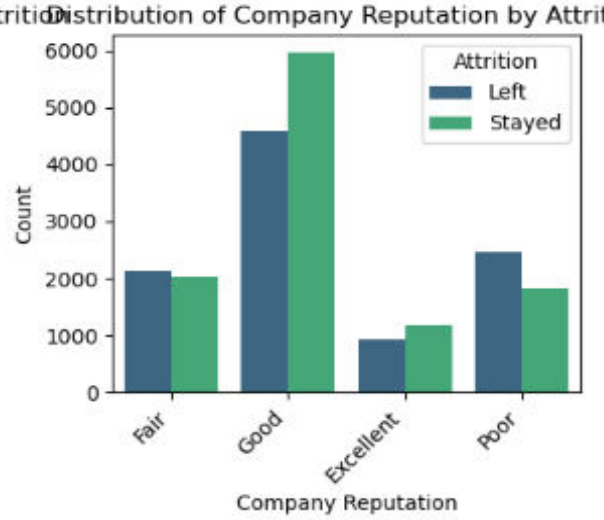
BIVARIATE ANALYSIS(CATEGORICAL)

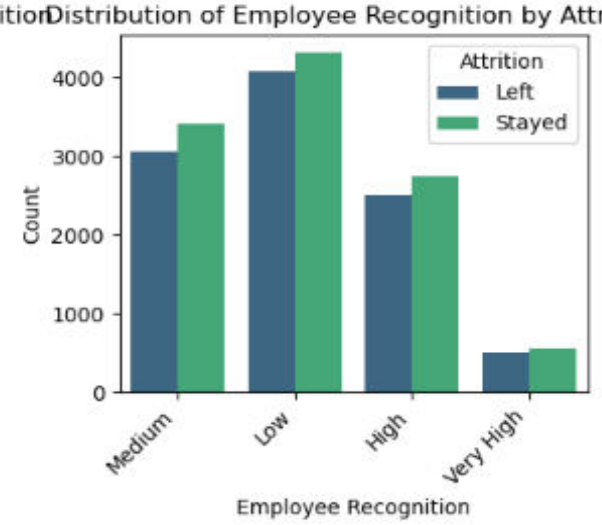












INSIGHTS(BIVARIATE ANALYSIS)

- **Females** have higher attrition rate in comparison to males.
- **Poor work-life balance** correlates with higher attrition.
- Employees with **no remote work option** showed **slightly higher attrition rates.**
- Employees with **"Below Average" performance ratings** were more likely to leave.
- Employees doing **overtime** are more likely to leave – burnout can be a possible reason.
- Employees having '**Marital Status**' as **Single** are more likely to leave.
- **Entry level** employees having high attrition rate.
- **Small size** company have high attrition rate in comparison to medium and large size company.
- **Company** having **Poor Reputation** have high attrition rates.

Model building

- After feature selection with rfe we built our GLM model.
- p-value are within range.
- vif are within range.

	Features	VIF
0	Gender_Male	1.84
10	Job Level_Mid	1.65
9	Marital Status_Single	1.42
1	Work-Life Balance_Fair	1.41
7	Overtime_Yes	1.38
11	Job Level_Senior	1.32
14	Company Reputation_Poor	1.26
13	Company Reputation_Fair	1.25
4	Job Satisfaction_Very High	1.23
2	Work-Life Balance_Poor	1.18
12	Remote Work_Yes	1.18
5	Performance Rating_Below Average	1.15
3	Job Satisfaction_Low	1.12
6	Performance Rating_Low	1.05
8	Education Level_PhD	1.05

Generalized Linear Model Regression Results

Dep. Variable:	Attrition_Stayed	**No. Observations:**	49390
Model:	GLM	**Df Residuals:**	49374
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-25012.
Date:	Sun, 20 Apr 2025	**Deviance:**	50024.
Time:	04:18:39	**Pearson chi2:**	4.61e+04
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.3104
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2461	0.028	8.701	0.000	0.191	0.301
Gender_Male	0.5860	0.022	26.379	0.000	0.543	0.630
Work-Life Balance_Fair	-1.0822	0.025	-42.915	0.000	-1.132	-1.033
Work-Life Balance_Poor	-1.2256	0.034	-36.424	0.000	-1.292	-1.160
Job Satisfaction_Low	-0.5024	0.037	-13.498	0.000	-0.575	-0.429
Job Satisfaction_Very High	-0.4796	0.028	-17.313	0.000	-0.534	-0.425
Performance Rating_Below Average	-0.3309	0.031	-10.714	0.000	-0.391	-0.270
Performance Rating_Low	-0.6030	0.051	-11.758	0.000	-0.704	-0.503
Overtime_Yes	-0.3237	0.023	-13.806	0.000	-0.370	-0.278
Education Level_PhD	1.5124	0.055	27.346	0.000	1.404	1.621
Marital Status_Single	-1.7050	0.025	-69.085	0.000	-1.753	-1.657
Job Level_Mid	0.9704	0.024	40.063	0.000	0.923	1.018
Job Level_Senior	2.4894	0.034	72.335	0.000	2.422	2.557
Remote Work_Yes	1.7257	0.032	53.570	0.000	1.663	1.789
Company Reputation_Fair	-0.5438	0.029	-19.062	0.000	-0.600	-0.488
Company Reputation_Poor	-0.7361	0.028	-25.843	0.000	-0.792	-0.680

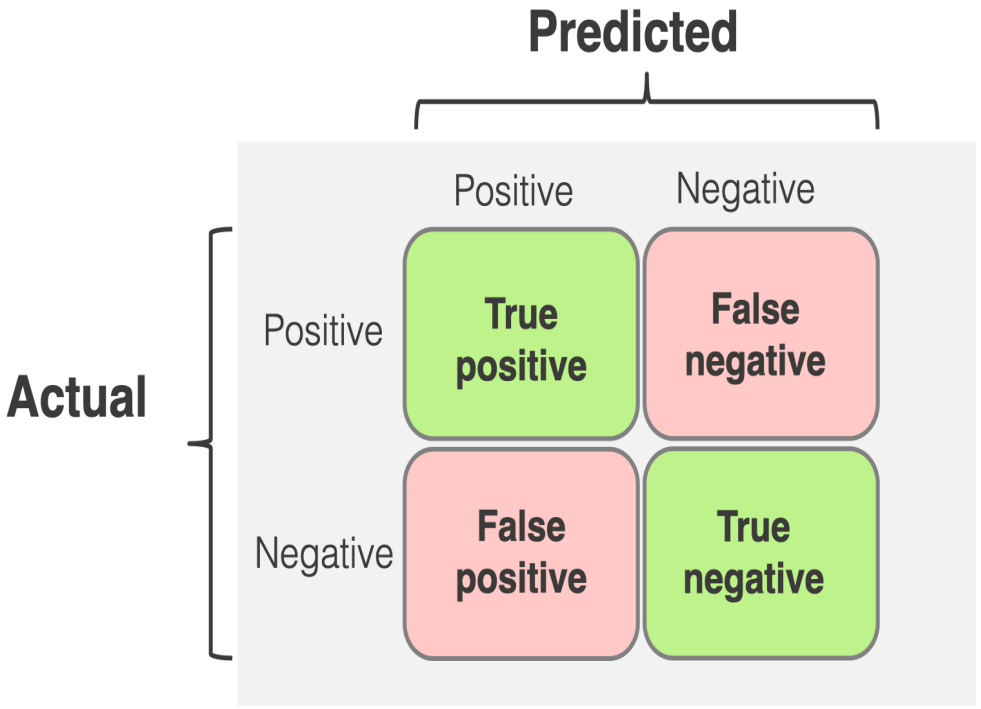
Model Performance

- First we create confusion matrix

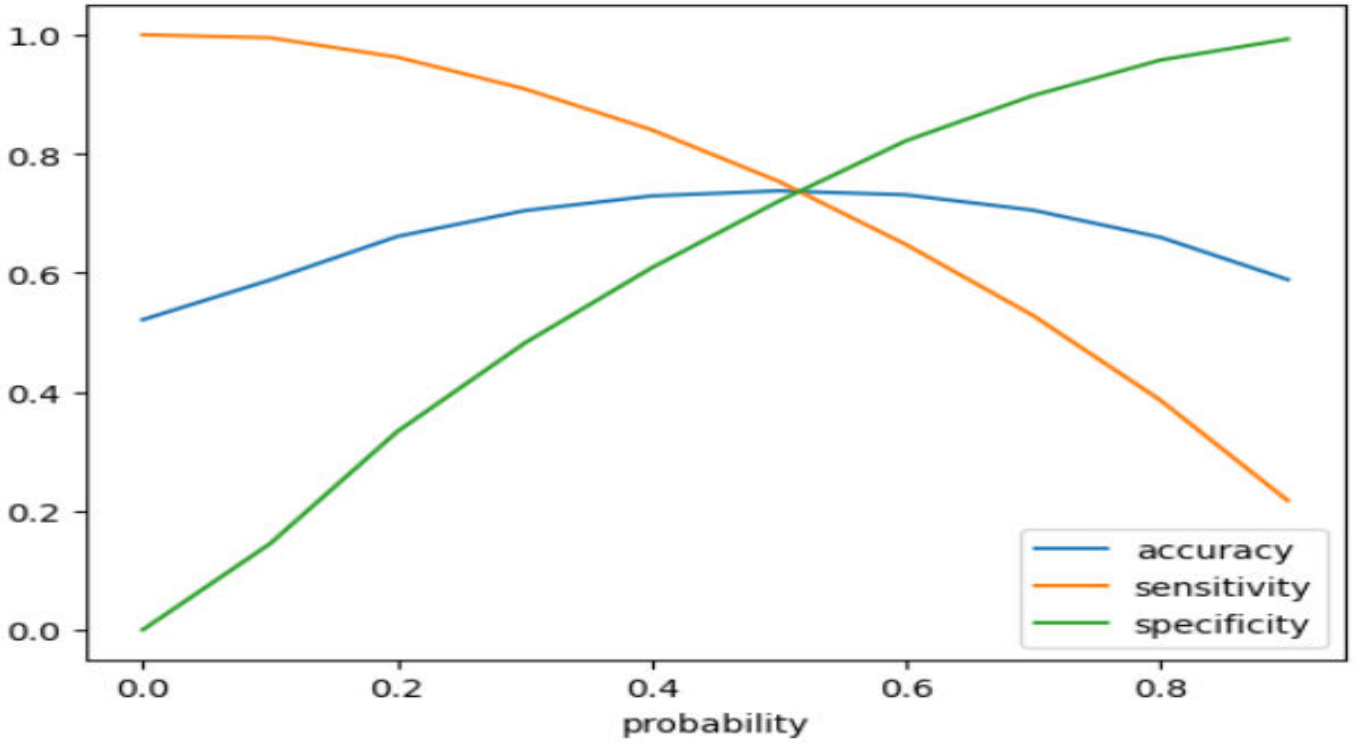
1. Accuracy = TP+TN/(TN+FP+FN+TP)
2. Sensitivity = TP/(TP+FN) = Recall
3. Specificity = TN/(FP+TN)
4. Precision = TP /(TP+FP)

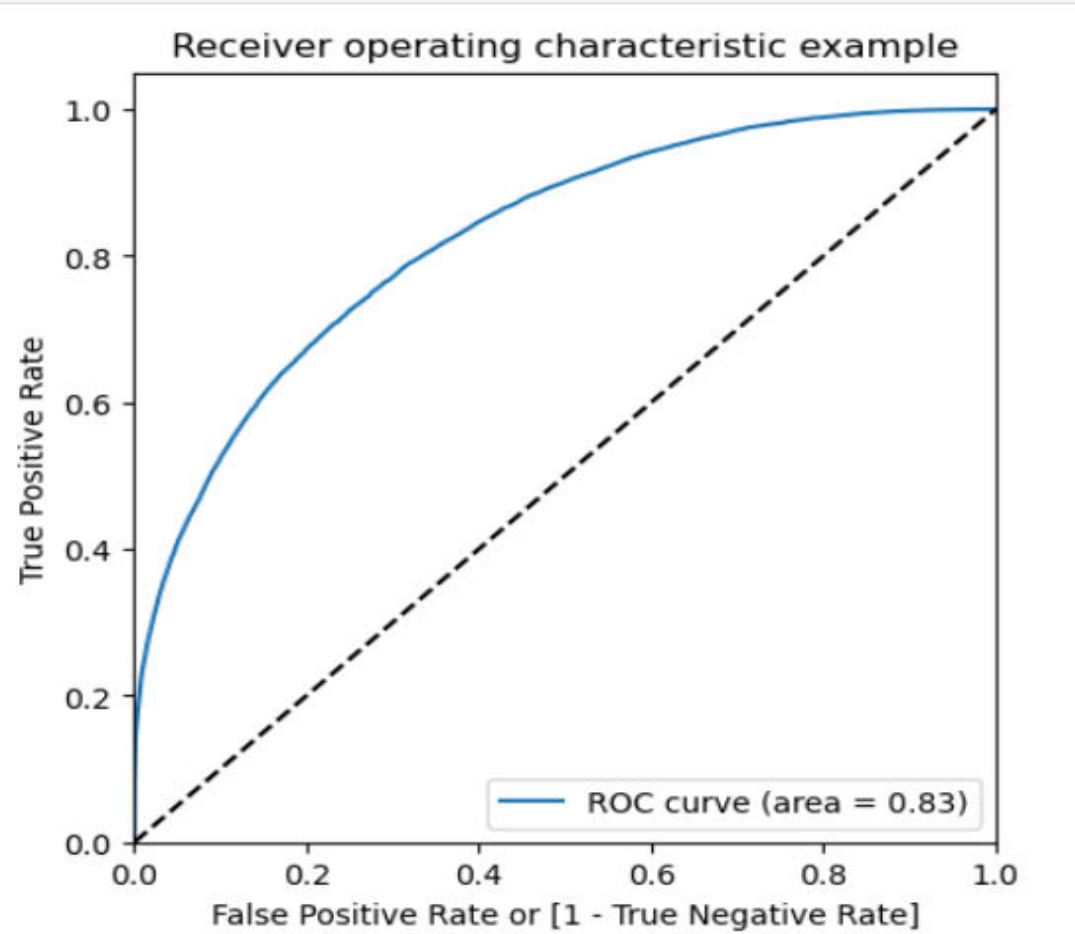
- For training data

1. Accuracy = 0.7376
2. Sensitivity=0.7535 = Recall
3. Specificity = 0.7204
4. Precision = 0.7457



ROC Curve and optimal cut-off

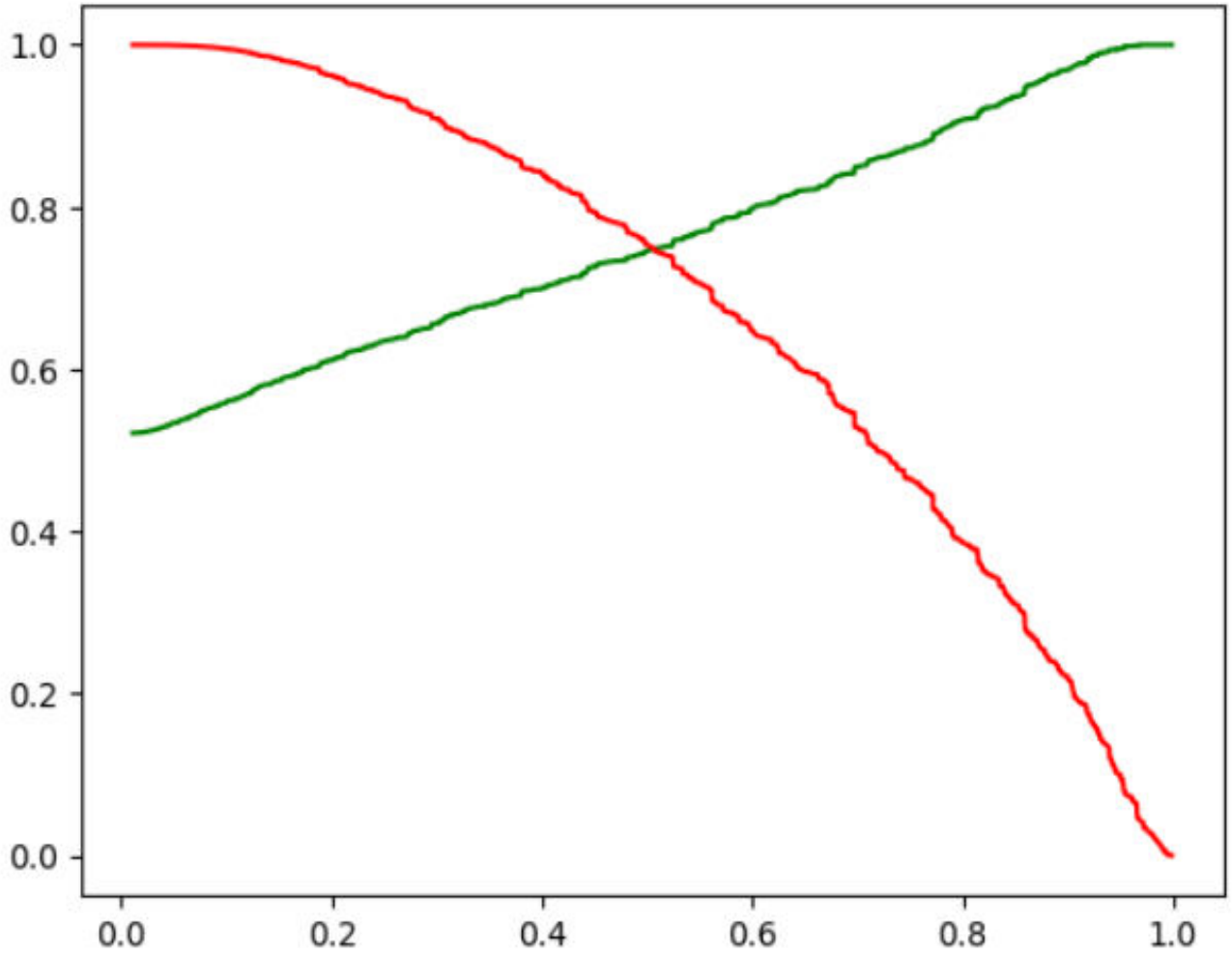




- From the ROC curve **0.5** is identified as the **optimum threshold** for classifying attrition. This value balances the **true positive rate** (sensitivity) and the **false positive rate**, ensuring effective decision-making for employee retention predictions.

Precision-Recall Curve

- **Precision:** Measures how many of the predicted "attrition" cases are actually correct. If it's **high**, the model makes fewer false positives.
- **Recall:** Measures how well the model captures all actual attrition cases. If it's **low**, the model misses real exits.
- Optimal-cutoff is 0.5.



Model Performance (validation)

1. Accuracy = 0.7396
2. Sensitivity= 0.7535
3. Specificity= 0.7204
4. Precision= 0.7457

- The Logistic Regression model for employee attrition prediction showed **73.9% accuracy**, correctly classifying employees staying or leaving. It achieved **75.3% recall**, effectively identifying potential attrition cases, and **74.5% precision**, ensuring valid predictions. The model had **72% specificity**, meaning it correctly identified employees who would stay

Conclusion

Factors Affecting Attrition

- **Poor work-life balance correlates with higher attrition**—indicating burnout as a potential issue.
- **Income Impact:** Employees with **lower salaries tend to leave more**
- Employees with **"Below Average" performance ratings** were more likely to leave.
- No promotions demotivates employees and leads to attrition.
- Making employee work overtime leads to high attrition.

Recommendations

- **Increase salary competitiveness** for employees in high-attrition brackets.
- **Enhance promotion policies** for long-serving employees to improve retention.
- **Improve work-life balance initiatives** to reduce burnout and disengagement.
- **Introduce flexible remote work options** where feasible.
- **Develop leadership programs** to engage employees looking for career growth.