Predicting Employee Retention

The objective of this assignment is to develop a Logistic Regression model. We will be using this model to analyse and predict binary outcomes based on the input data.

Problem Statement

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.

Assumptions

- Dropped the rows with null values as the percentage of null values was very less.
- Dropped the redundant columns which will not help much in model building.
- Cleaned the data to prepare for analysis and model creation.
- Perform EDA univariate analysis and Bi-variate analysis on Train_data and Validation_data.

Approach

1. Data Cleaning

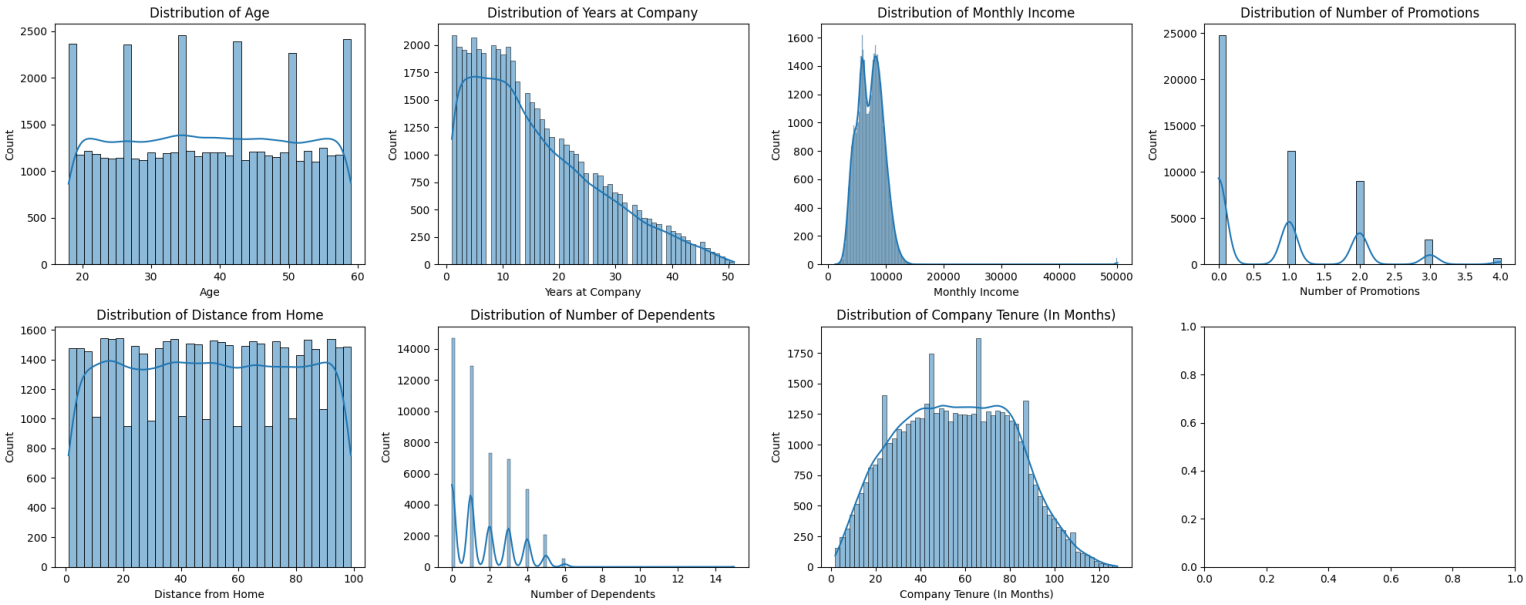
- Quick check was done on % of null value, and we have dropped all the rows with null values as the percentage of null values was very less around 3-4%.
- We checked all the categorical columns and checked for redundant values, and we have dropped “Employee ID” as this would not contribute in model building and “Leadership Opportunities” as data is very skewed and will not add much value.

2. Train-Validation Split

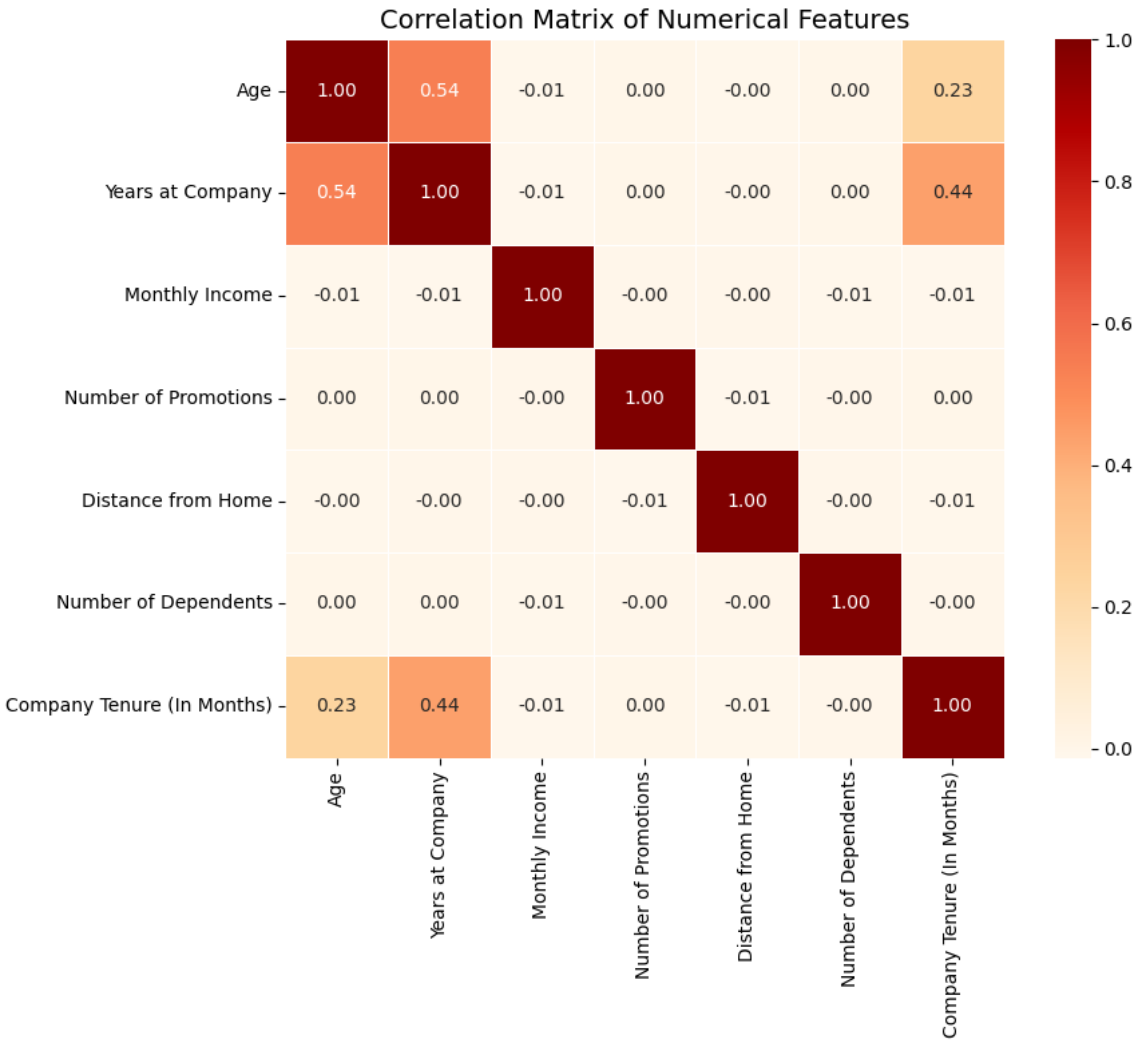
- The split was done at 70% and 30% for train and test data respectively.

3. EDA

- Checked distribution for all the numeric variables

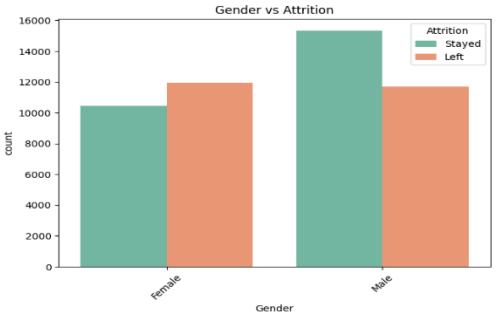


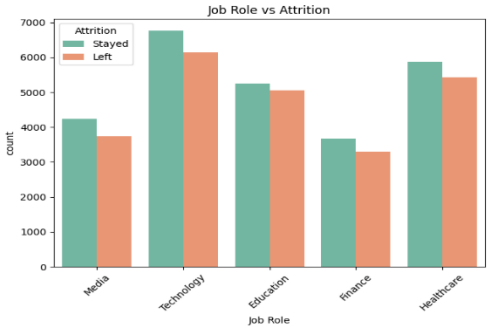
- Checked the correlation matrix for the numeric variables

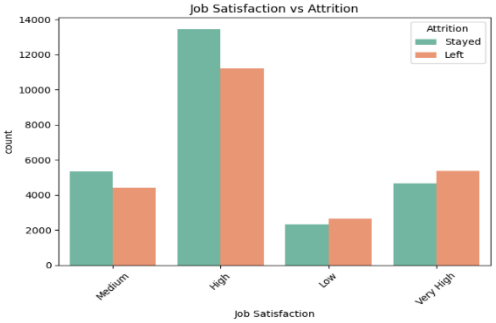


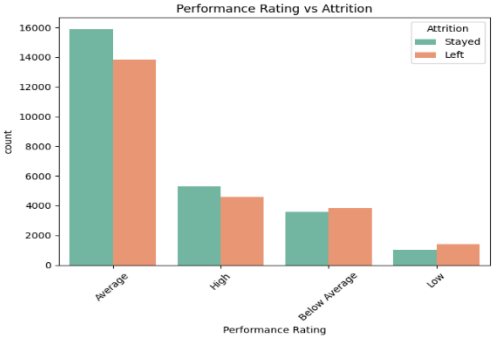
- Checked the class balance for the dependent variable

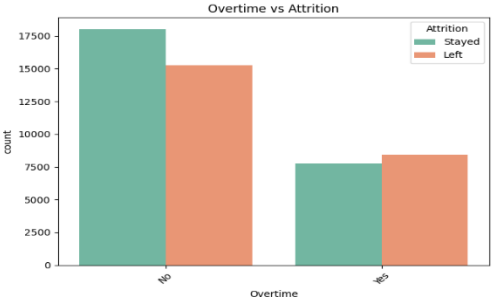
- Performed the bivariate analysis between all the categories columns and target variables (Attrition)

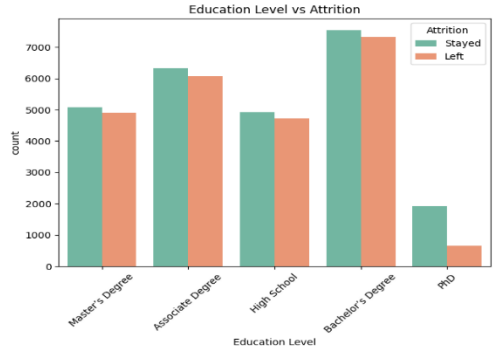


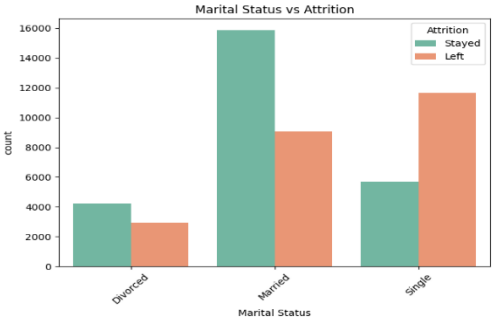


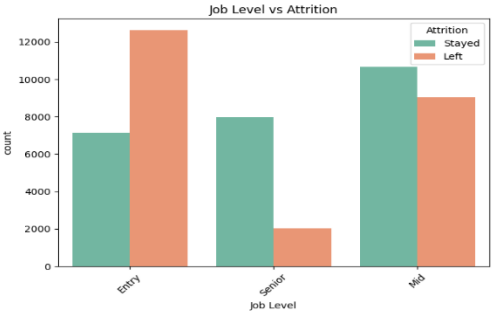


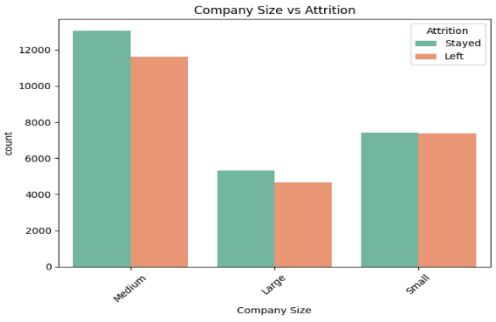


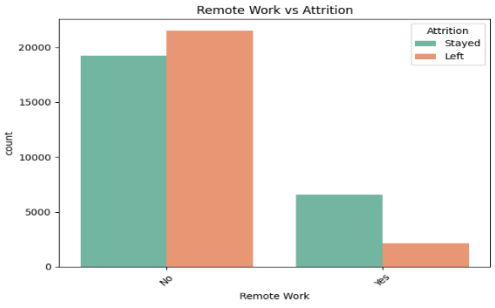


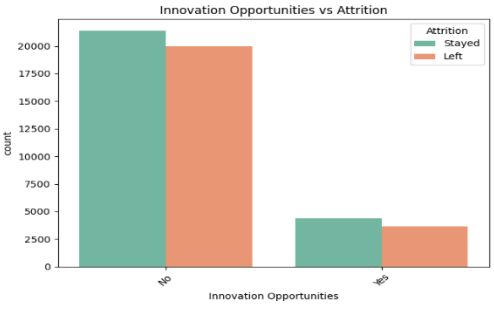


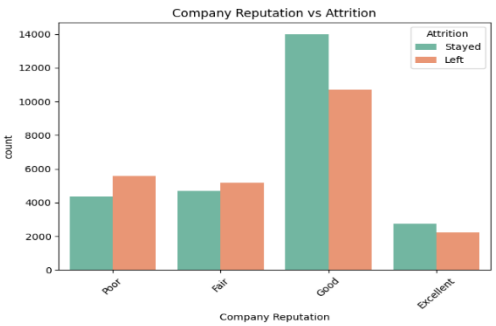


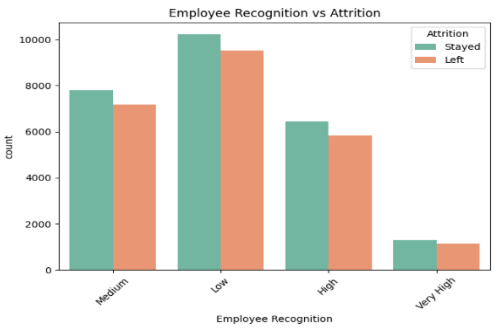












4. Feature Engineering

- Created dummy variables for all the categorical columns and target variable.
- Scaled all the numerical variables using Standard Scaler.

5. Model Building

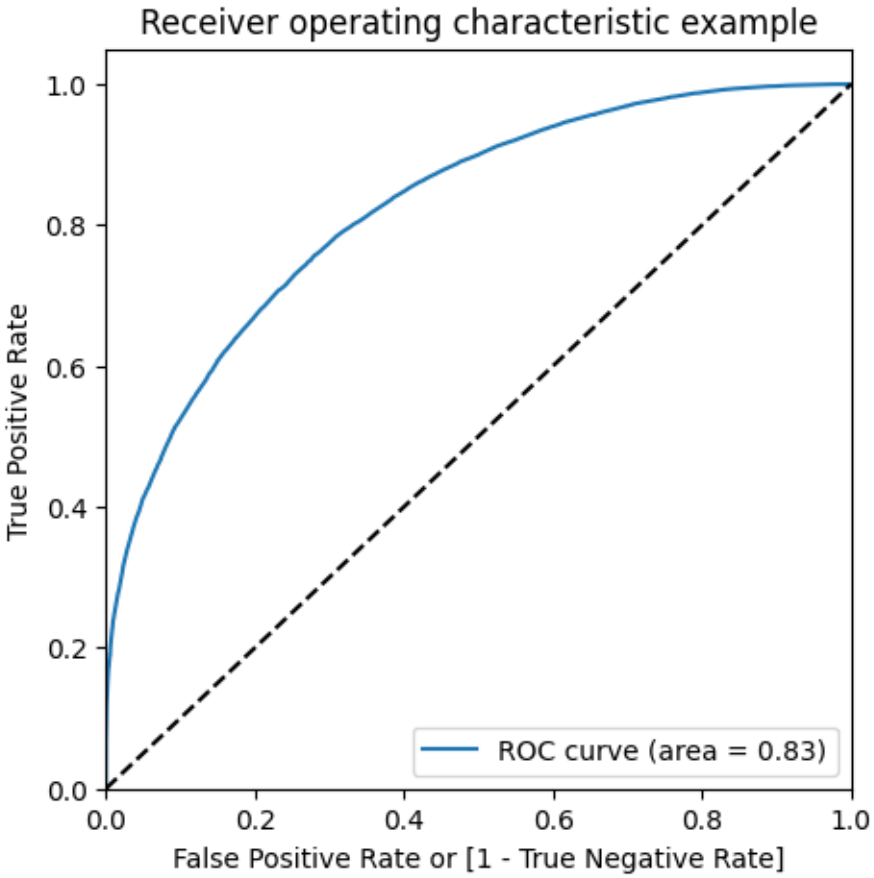
- RFE was used for feature selection.
- Then RFE was done to attain the top 15 relevant variables.
- Built the model using Generalized Linear Model (GLM) for prediction

Generalized Linear Model Regression Results

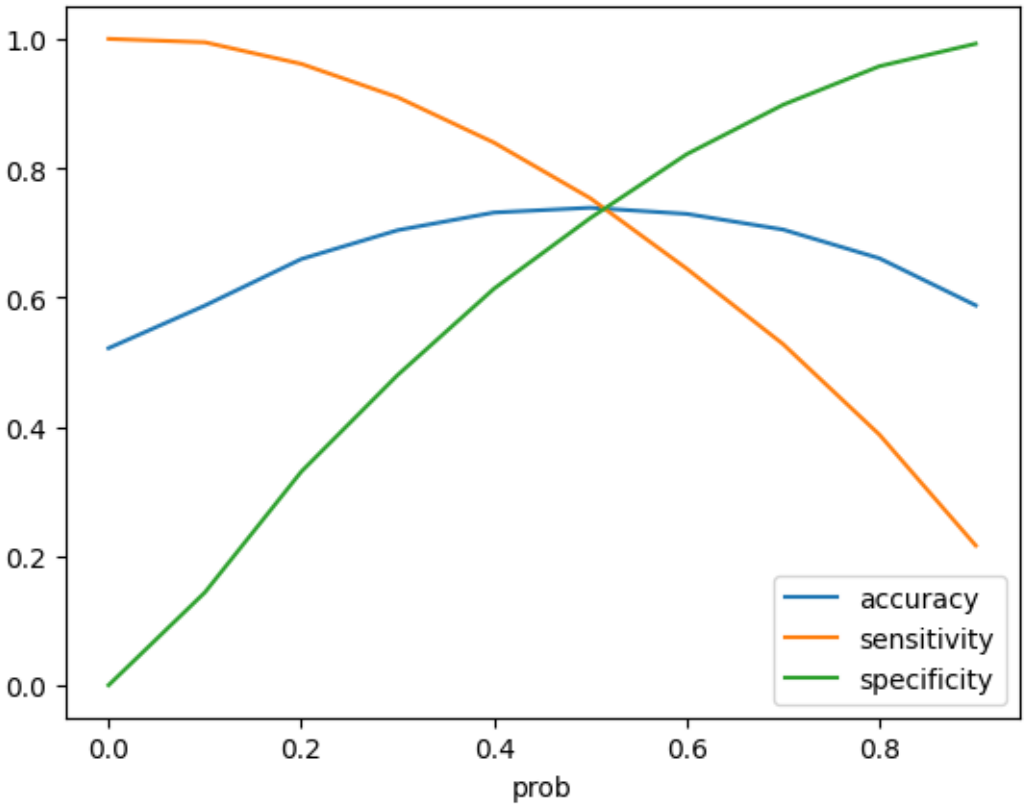
Dep. Variable:	Attrition_Stayed	**No. Observations:**	49444
Model:	GLM	**Df Residuals:**	49428
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-25072.
Date:	Sun, 20 Apr 2025	**Deviance:**	50143.
Time:	20:21:47	**Pearson chi2:**	4.64e+04
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.3095
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2356	0.028	8.336	0.000	0.180	0.291
Gender_Male	0.5738	0.022	25.875	0.000	0.530	0.617
Work-Life Balance_Fair	-1.0572	0.025	-41.957	0.000	-1.107	-1.008
Work-Life Balance_Poor	-1.2489	0.034	-37.262	0.000	-1.315	-1.183
Job Satisfaction_Low	-0.4914	0.037	-13.282	0.000	-0.564	-0.419
Job Satisfaction_Very High	-0.4829	0.028	-17.462	0.000	-0.537	-0.429
Performance Rating_Below Average	-0.3079	0.031	-10.012	0.000	-0.368	-0.248
Performance Rating_Low	-0.5673	0.051	-11.058	0.000	-0.668	-0.467
Overtime_Yes	-0.3286	0.023	-14.053	0.000	-0.374	-0.283
Education Level_PhD	1.4826	0.055	27.084	0.000	1.375	1.590
Marital Status_Single	-1.6887	0.025	-68.644	0.000	-1.737	-1.640
Job Level_Mid	0.9611	0.024	39.780	0.000	0.914	1.008
Job Level_Senior	2.5326	0.035	73.125	0.000	2.465	2.600
Remote Work_Yes	1.7154	0.032	53.379	0.000	1.652	1.778
Company Reputation_Fair	-0.5296	0.028	-18.633	0.000	-0.585	-0.474
Company Reputation_Poor	-0.7359	0.029	-25.758	0.000	-0.792	-0.680

- Plotted the ROC curve to check area under the curve. (area = 0.83)



- Plotted the accuracy, sensitivity, and specificity at different values of probability cutoffs to find the optimal cutoff, which comes out to be at **0.5**



6. Prediction and Model Evaluation

- Evaluation for training data:
  a. Sensitivity : 75.33%
  b. Specificity : 72.28%
  c. Precision : 74.75%
  d. Recall : 75.33%

- Evaluation for validation data:
  a. Sensitivity : 74.65%
  b. Specificity : 72.40%
  c. Precision : 74.59%
  d. Recall : 74.65%

Conclusion

Through the model built we can see what features effecting the highest for retention of employees.

1. Remote work options significantly increase retention likelihood, suggesting flexible work arrangements can strengthen employee loyalty.
2. Higher job levels, especially Senior roles, are strong indicators of retention, highlighting the value of career growth opportunities.
3. Poor or fair work-life balance greatly reduces the chances of retention, underlining the need for better work-life integration policies.
4. Single employees are less likely to stay compared to their married counterparts, indicating the potential benefit of inclusive engagement initiatives for diverse life situations.
5. Employees with PhDs are more likely to remain, suggesting investment in advanced learning and development can support long-term retention.
6. Working overtime is negatively associated with retention, pointing to the risk of burnout and the importance of workload management.
7. Low job satisfaction and performance ratings decrease the likelihood of staying, emphasizing the need to improve workplace satisfaction and recognize good performance.