Predicting_Employee_Retention_Parth_Kalpesh_Radhika

Objective

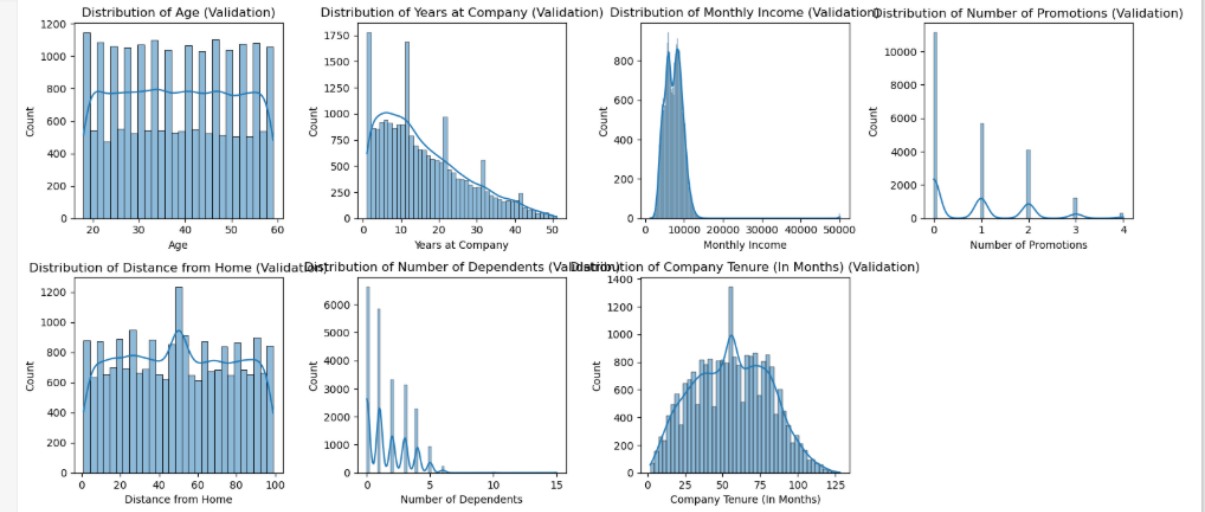
The objective of this assignment is to develop a Logistic Regression model. You will be using this model to analyse and predict binary outcomes based on the input data. This assignment aims to enhance understanding of logistic regression, including its assumptions, implementation, and evaluation, to effectively classify and interpret data.

Conclusion

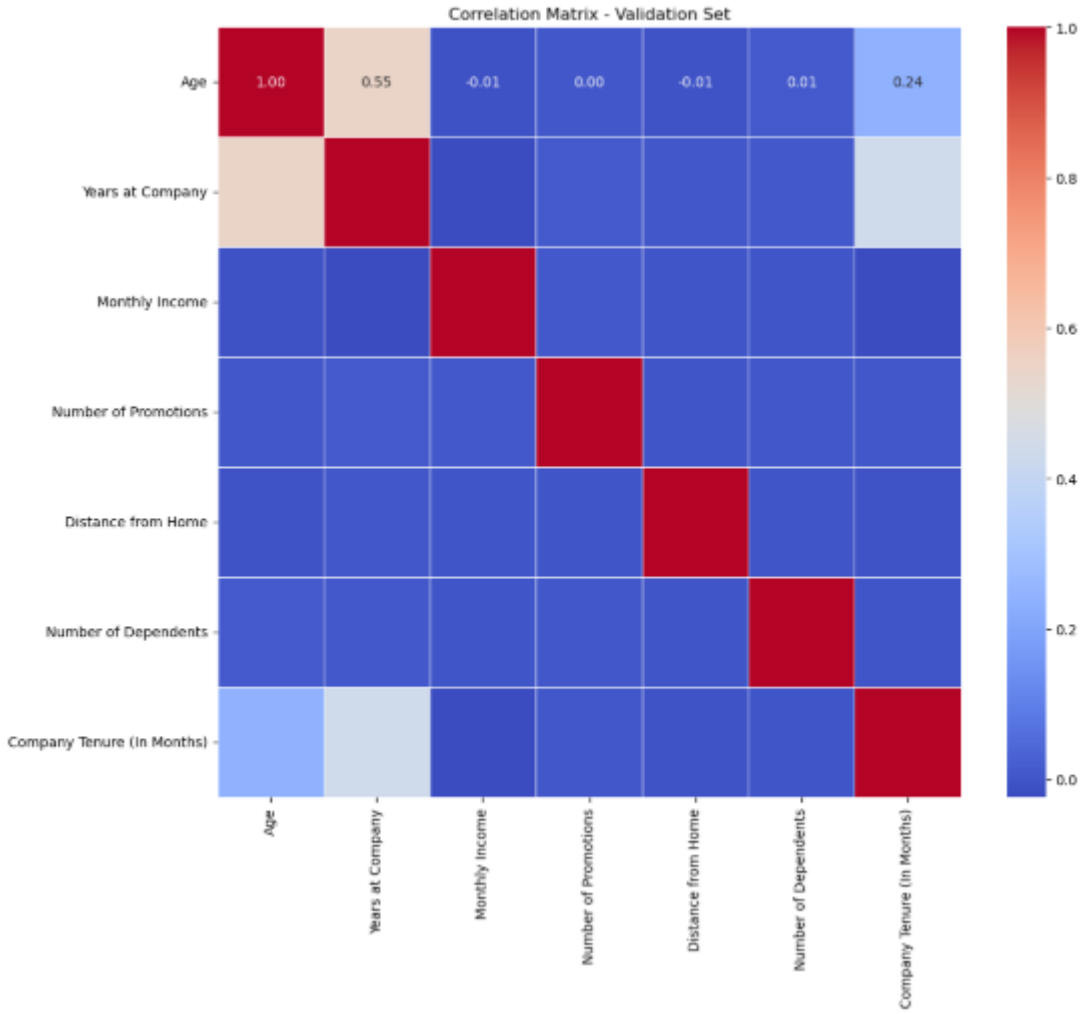
In this assignment, we developed a logistic regression model to predict employee retention based on various factors such as job satisfaction, performance, income, recognition, and more.

We began by thoroughly 1) Loading the data, cleaning the data—handling missing values, standardizing categorical entries, and removing redundancies. 2) After splitting the dataset into training and validation sets (in a 70:30 ratio), 3) we performed exploratory data analysis (EDA) on the training data to understand feature distributions, class imbalance, and relationships between variables and attrition. 6) For feature engineering, we applied Dummy to categorical variables and scaled numerical features to ensure consistency. 7) Using Recursive Feature Elimination (RFE), we selected the most relevant features for the model. We then built a logistic regression model and evaluated its performance. To find the optimal threshold for classification, 8) we analysed the trade-off between sensitivity and specificity and selected a cut-off of **0.50**, which gave us a balanced performance with around **71% accuracy**.

From the **distribution plots**, we observed that some features like **Distance from Home**, **Monthly Income**, and **Years at Company** showed right-skewed distributions, indicating the presence of outliers or rare patterns that needed standardization.



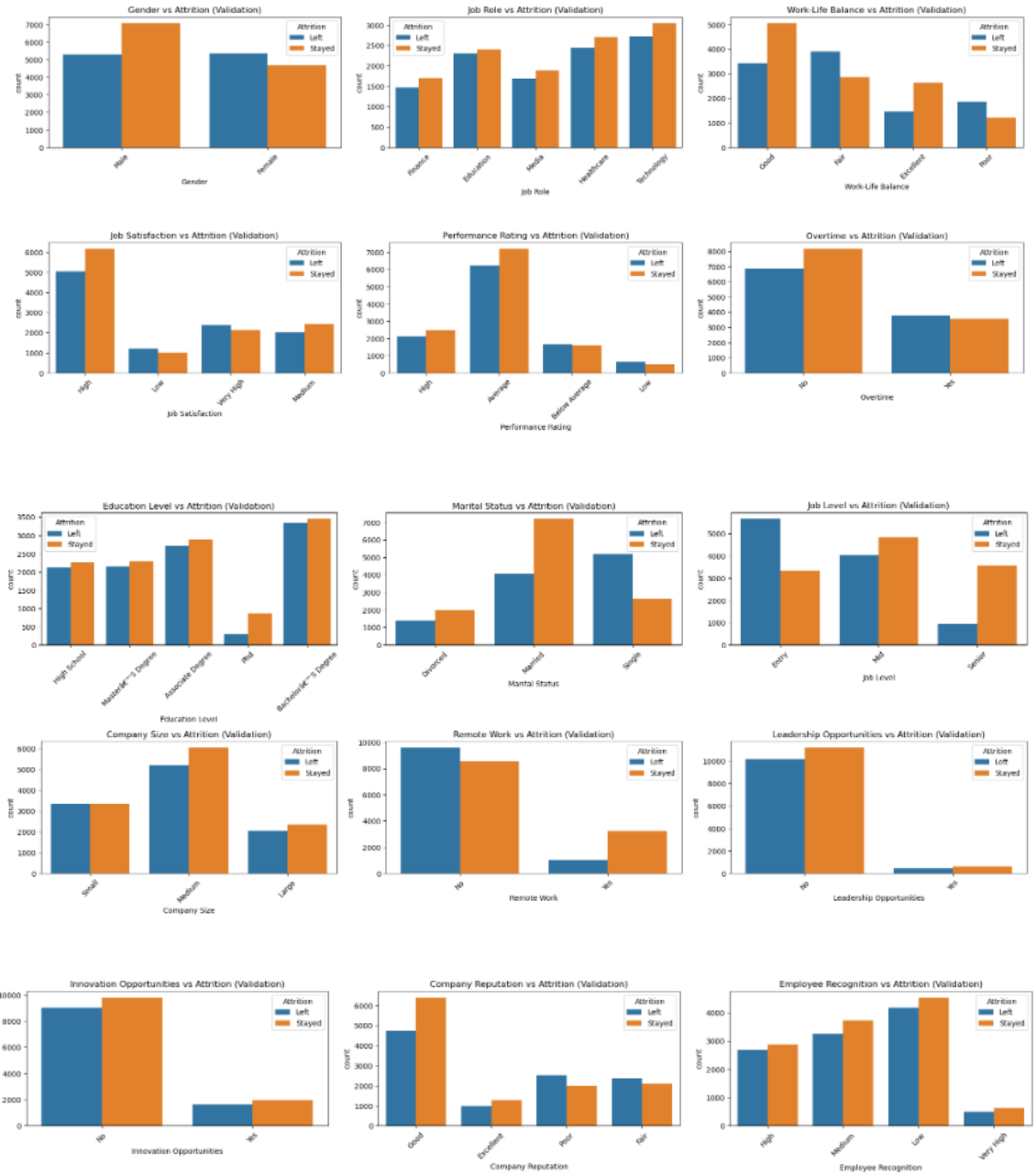
The **correlation heatmap** helped us detect multicollinearity among numerical features. For example, **Job Level** and **Monthly Income** were strongly correlated, which aligns with business intuition.



The **class distribution plot** revealed a mild imbalance, with a smaller percentage of employees leaving the company. This was important to consider while selecting evaluation metrics.

From the **bivariate analysis** using count plots, we found that:

- Employees with **low work-life balance**, **low recognition**, or **very low job satisfaction** had much higher attrition.
- **Overtime hours** and **lack of leadership or innovation opportunities** also showed a strong correlation with employee exit.


Gender vs Attrition (Validation)
Attrition
Left
Stayed
count
Male
Female
Gender
Job Role vs Attrition (Validation)
Finance
Education
Media
Healthcare
Technology
Job Role
Work-Life Balance vs Attrition (Validation)
Good
Fair
Excellent
Poor
Work-Life Balance
Job Satisfaction vs Attrition (Validation)
High
Low
Very High
Medium
Job Satisfaction
Performance Rating vs Attrition (Validation)
High
Average
Below Average
Low
Performance Rating
Overtime vs Attrition (Validation)
No
Yes
Overtime
Education Level vs Attrition (Validation)
High School
Master's Degree
Associate Degree
PhD
Bachelor's Degree
Education Level
Marital Status vs Attrition (Validation)
Divorced
Married
Single
Marital Status
Job Level vs Attrition (Validation)
Entry
Mid
Senior
Job Level
Company Size vs Attrition (Validation)
Small
Medium
Large
Company Size
Remote Work vs Attrition (Validation)
No
Yes
Remote Work
Leadership Opportunities vs Attrition (Validation)
No
Yes
Leadership Opportunities
Innovation Opportunities vs Attrition (Validation)
No
Yes
Innovation Opportunities
Company Reputation vs Attrition (Validation)
Good
Excellent
Poor
Fair
Company Reputation
Employee Recognition vs Attrition (Validation)
High
Medium
Low
Very High
Employee Recognition

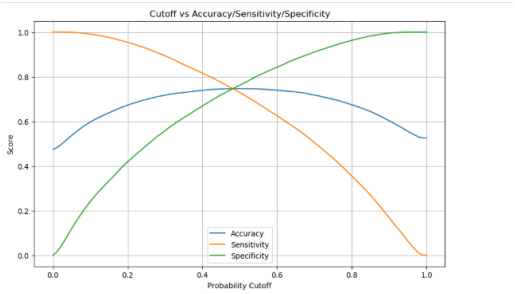
Finally, we evaluated the model on the validation set using metrics such as accuracy, confusion matrix, sensitivity, specificity, precision, and recall. The results confirmed that the model was reliable and interpretable for predicting employee retention.

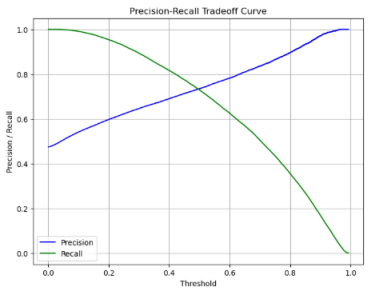
Sensitivity: 0.8925277620929795

Specificity: 0.5597516373224461

Precision: 0.6469304229195089

Recall: 0.8925277620929795





The optimal cut-off was found to be 0.50, which balances both sensitivity and specificity, achieving approximately 71% accuracy on the training data

The overall accuracy of the model is 71%, showing it works reasonably well.