Predicting Employee Retention

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 recognizes the value of proactively identifying employees likely to stay and understanding the factors contributing to their loyalty.

The goal of this assignment is to develop a logistic regression model to predict employee retention based on 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 overall workforce stability.

Methodology

The assignment followed a structured approach:

1. **Data Understanding**
   Reviewed a dataset with 24 features and over 74,610 records. Columns included demographic details, job-level metrics, and the target variable indicating retention.
2. **Data Cleaning**
   a. Addressed missing values.
   b. Identify and handle redundant values
   c. Categorical encoding and numerical scaling were applied. d. Checked for outliers and inconsistent data entries.
3. **Train-Validation Split**
   Split the data into training and validation into 70% train data and 30% validation datasets to prevent overfitting and ensure proper model evaluation.
4. **Exploratory Data Analysis (EDA)**
   a. Distribution plots for numerical features to identify skewness and outliers.
   b. Univariate analysis is conducted by visualising the distribution of all numerical columns.
   c. Correlation analysis is conducted by visualising a heat map of the correlation matrix to detect multicollinearity.
   d. Count plots to examine class imbalance.
   e. Bivariate analysis is performed by visualising the relationship between categorical columns and the target variable.
5. **Feature Engineering**
   a. Created dummy variables for independent and dependent columns(for training and test data) in both training and validation and interaction terms to improve model performance.
   b. **StandardScaler** from scikit-learn was used to numerical columns . It helps the logistic regression model converge faster and assign balanced weights to features.
6. **Model Building**
   a. Selected the important feature using **Recursive Feature Elimination** (RFE).
   b. Used **logistic regression** as the classification algorithm. Assessed multicollinearity using p-values and Variance Inflation Factors (VIF), then generated predictions and evaluated the model's effectiveness using appropriate performance metrics.

c. Fit the model on training data and adjusted hyperparameters through validation testing. Evaluate the performance of the model based on the predictions made on the training set.

d. Cutoff optimization plot is used to Check **sensitivity and specificity** tradeoff to find the optimal cutoff point at various probability thresholds.

7. **Prediction and Model Evaluation**
   Evaluated the model using accuracy, precision, recall, F1-score, and ROC-AUC to determine the classification performance. Evaluated the model's performance correctly using the given evaluation metrics.

Techniques Used

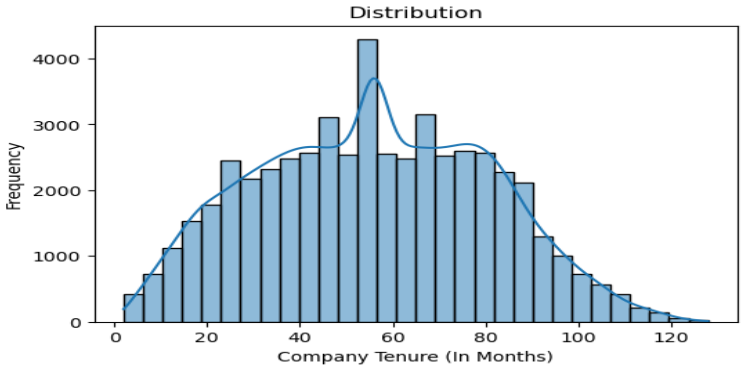
a) Logistic Regression
b) Feature Scaling and One-Hot Encoding
c) Data Visualization with cutoff optimization plot , Seaborn and Matplotlib
d) Evaluation Metrics: Accuracy, Precision, Recall, F1 Score, ROC-AUC

Assumptions

a) Redundant columns were droped like Employee Recognition ,Innovation Opportunities
b) Dummies were created
c) The model showed that while a 0.5 cutoff balances both metrics (**sensitivity and specificity)**.

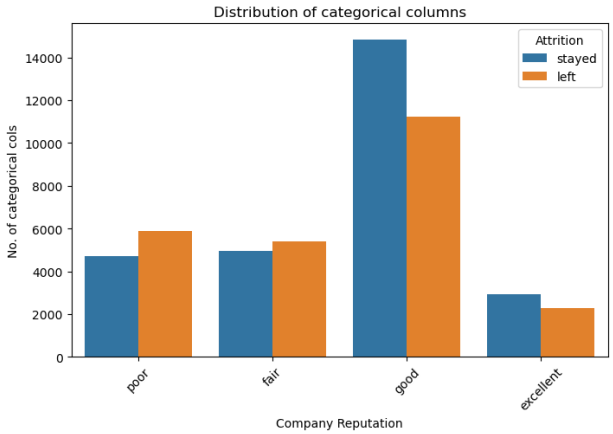
Visualisation

a) Plot distribution of numerical columns

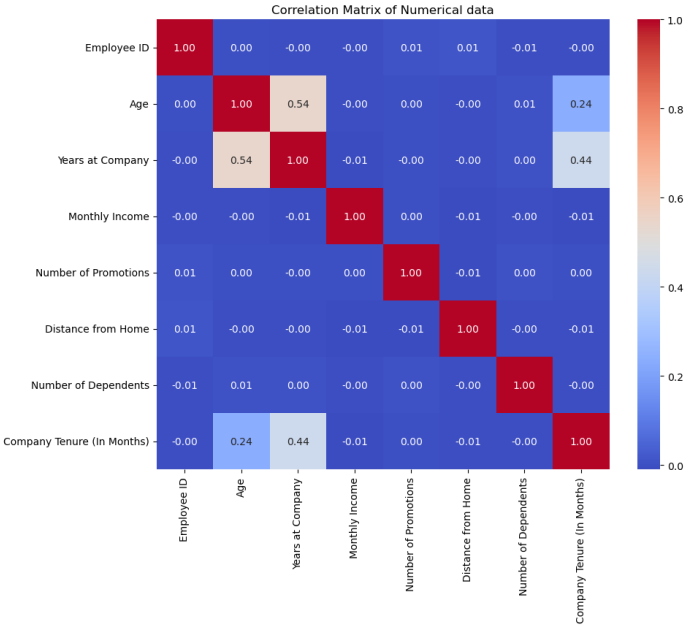


- Plot a graph on numerical columns given in the data(Employee ID', 'Age', 'Years at Company', 'Monthly Income''Number of Promotions', 'Distance from Home', 'Number of Dependents','Company Tenure (In Months)'.
-

b) Graph plotted on categorical columns, we plot with legend attrition.



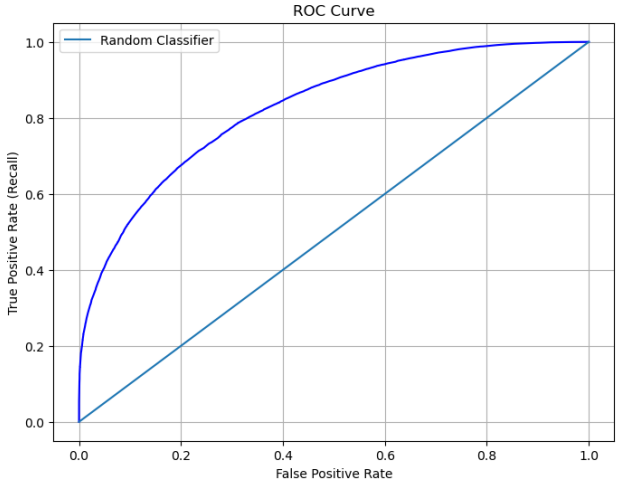
c) EDA- correlation among different numerical variables



- Age and Years at Company as well as Years at Company and Company Tenure show moderate positive correlation, which is expected due to the natural relationship between employee age, experience, and tenure. This low multicollinearity supports the use of these variables in logistic regression.

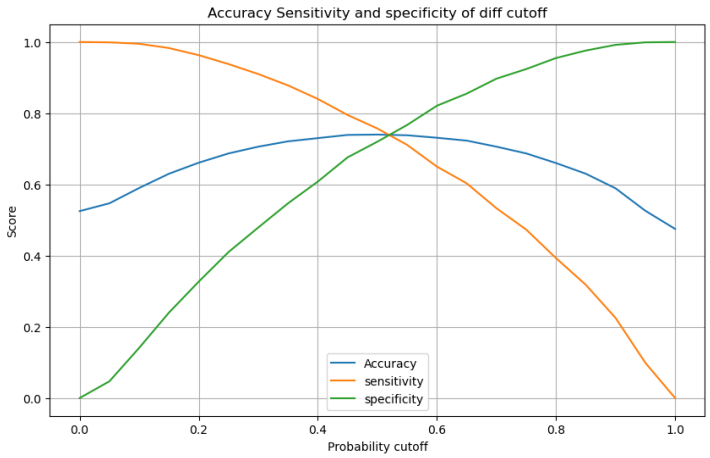
d) In EDA graph plotted univariate analysis for numerical columns for training data and validation data.

e) Roc curve



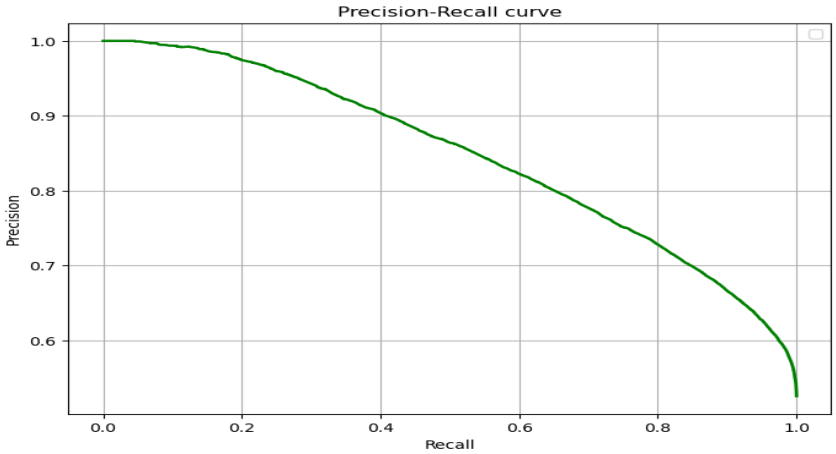
- Roc curve visualization for actual and predicted probability,Blue curve shows trade of between TPR and FPR. Roc curve is top left corner so its better model.

f) Plot accuracy, sensitivity, and specificity at different values of probability cutoffs



- In accuracy, sensitivity and specificity cutoff graph shows the intersaction point is approx(0.5) which we can consider balance threshold.

g) Plot precision-recall curve



- In precision recall curve we get high precision and low recall which is good behavior model,and no sharp fall so its stable.

Key Insights

The logistic regression model revealed several important predictors of employee retention:

a) **Job Satisfaction** and **Work-Life Balance** emerged as strong positive influencers—employees with higher scores in these areas are significantly more likely to stay.
b) **Overtime** and a **lack of promotions** were associated with increased attrition risk.
c) The class distribution of the target variable (Attrition) is relatively balanced, with a slight majority of employees who stayed. This balance reduces the risk of model bias toward the majority class .
d) The model showed that while a 0.5 cutoff balances both metrics , lower the thresholds increases the sensitivity making it more effective at indentify the employees at risk of leaving.
e) Employees with **high satisfaction** had the lowest attrition rates, whereas those with **low or medium satisfaction** were more likely to leave, but its not only the sole factor.
f) **Age and Years at Company** as well as **Years at Company and Company Tenure** show moderate positive correlation, which is expected due to the natural relationship between employee age,

experience, and tenure. This low multicollinearity supports the use of these variables in logistic regression.

g) The ROC curve is well above the diagonal line, which indicates that **the model is performing significantly better than random.** This suggests that the model reliably distinguishes between employees who are likely to stay versus those at risk of leaving

h) **Performance Rating** also played a crucial role; high-performing employees are more likely to be retained.

These insights suggest that HR can **proactively intervene** for employees flagged as high-risk for attrition. By focusing on improving satisfaction, recognizing performance, and managing workload, the company can build a more stable and satisfied workforce.