**Problem Statement**

A mid-sized technology company is seeking to improve its **employee retention strategy**. Instead of reacting to employee turnover, they want to proactively **predict which employees are likely to stay**, using historical data. The goal is to use **logistic regression** to model this binary classification problem and identify key drivers of retention.

**Methodology & Techniques Used**

**1. Exploratory Data Analysis (EDA)**

* Initial examination of variable distributions, null values, and outliers.
* Categorical features were explored using value counts and visualizations.
* Numerical features were analyzed using histograms and correlation matrices.

**2. Data Preprocessing**

* Categorical variables were encoded using **dummy variables**.
* Outliers were handled, and features were standardized if needed.
* The dataset was split into **training and test sets** using train\_test\_split.

**3. Model Building**

* A **Logistic Regression model** was built using statsmodels for interpretability.
* Model summary was used to assess statistical significance of features (p-values).
* Important features were retained based on VIF and multicollinearity checks.

**4. Model Evaluation**

* Predicted probabilities were generated, and multiple **cutoff thresholds** were tested.
* Metrics used:
  + **Accuracy**
  + **Sensitivity (Recall for 1s)**
  + **Specificity (Recall for 0s)**
  + **Precision**
  + **F1 Score**
  + **ROC-AUC score**
* Optimal threshold was chosen using:
  + **Youden’s Index** on ROC Curve
  + **Maximum F1 Score** on Precision-Recall Curve

**5. Visualization**

* Plots included:
  + Precision & Recall vs Threshold
  + ROC Curve
  + Confusion Matrix at optimal cutoff
  + Accuracy/Sensitivity/Specificity tradeoff curve

**🔍 Key Insights**

1. **Important Predictors**:
   * Features like **Job Level**, **Remote Work**, **Work-Life Balance**, **Number of Promotions** and **Tenure** played significant roles.
   * Employees with higher satisfaction and performance scores were more likely to stay.
2. **Model Performance**:
   * The logistic regression model showed decent discriminatory power with a good ROC-AUC score.
   * The optimal cutoff was selected to balance **sensitivity and specificity**, crucial for retention strategies.
3. **Business Implications**:
   * The company can focus retention efforts on employees predicted as high risk for leaving.
   * Early interventions can be planned for low satisfaction/performance scores.
   * Provides HR with a data-backed tool for workforce planning.

**Conclusion**

* The logistic regression model successfully meets the objective of predicting employee retention.
* It provides interpretable insights and performs well on relevant classification metrics.
* This model can be integrated into HR dashboards to assist with **retention analytics**.