PREDICTING EMPLOYEE ATTRITION USING PYTHON-BASED DATA MINING FOR A U.S. TECHNOLOGY COMPANY

1. Background

A U.S.-based mid-sized technology company with over 1,000 employees experienced rising voluntary attrition, especially among experienced developers and product managers. Exit interviews highlighted dissatisfaction with growth and compensation, but HR lacked a data-driven method to flag at-risk employees early.

We were contracted to perform an end-to-end attrition risk analysis using Python. The goal was to build a predictive model using HR data, identify risk drivers, and output a segment-wise attrition score to guide retention strategies.

2. Objective

- To apply classification algorithms in Python to predict which employees are at risk of voluntary attrition
- To identify and rank the key features driving churn
- To deliver HR-friendly risk scores and actionable retention insights
- To support policy redesign around internal mobility, engagement, and performance feedback

3. Data Used

Source: Internal HRIS, employee surveys, and performance systems

Dataset Details:

- 1,032 active employees
- Historical attrition records from the past 2 years
- Fields included:
    - Employee_ID, Department, Tenure_Years, Last_Promotion_Years, Salary_Band, Performance_Rating, Training_Hours_Last_Year, Manager_Turnover, Remote_Status, Attrition (target)

4. Methodology

4.1 Data Preprocessing

- Cleaned missing values (e.g., filled performance gaps using team average)
- Normalized numeric fields like training hours and tenure
- One-hot encoded categorical features (Department, Remote_Status)
- Addressed class imbalance using **SMOTE** (Synthetic Minority Oversampling)

4.2 Modeling Approach

- Tested three models:
    - Logistic Regression
    - Random Forest
    - XGBoost
- Model evaluation using:
    - Precision, Recall, F1 Score, ROC AUC
    - 10-Fold Cross-Validation
    - SHAP (SHapley Additive exPlanations) for feature importance interpretation

5. Mining Results

Model	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.68	0.60	0.63	0.77
Random Forest	0.81	0.73	0.77	0.89
XGBoost	0.84	0.79	0.81	0.91

- **XGBoost** provided the highest overall predictive power
- Top predictors of attrition:
    - Last_Promotion_Years
    - Training_Hours_Last_Year
    - Salary_Band
    - Manager_Turnover

- Remote_Status (remote workers had slightly higher retention)

6. Strategic Insights

- Employees without promotion for >2 years were **2.4× more likely** to leave
- Undertrained staff (<10 training hours/year) had **higher churn risk**, regardless of performance
- Departments with high manager turnover showed consistent churn elevation
- Suggested creating **"Flight Risk" dashboard** for HRBPs to monitor key segments

7. Reporting Output

- **Python Notebook**:
  - Model pipeline for data cleaning, training, prediction, and evaluation
  - SHAP value plots to explain each prediction
  - Segment-based CSV outputs with attrition risk scores
- **PDF Report (19 pages)**:
  - Model summary and ROC curves
  - Feature impact heatmaps
  - Strategic HR recommendations (e.g., targeted promotions, retention bonus triggers)
- **Excel Dashboard**:
  - Employee-level attrition score (0–1)
  - Filters by department, tenure, training status
  - "Immediate Action" tag for top 10% high-risk profiles

8. Business Impact

- Within 3 months:
  - Voluntary attrition dropped from 7.9% to **6.7%**
  - High-risk list enabled **proactive interventions** (e.g., career coaching, internal transfers)

- Model now integrated into quarterly HR business reviews
- HR team trained to update the model monthly using rolling employee data