🔍 Predicting Employee Attrition with Machine Learning

"I built a classification model to identify drivers of employee attrition using IBM's fictional HR dataset."

Role: Data Analyst | Tools: Python, Scikit-Learn, XGBoost, Matplotlib

Business Problem

💼 Why This Matters

Problem statement:

- Employee attrition increases recruitment costs and disrupts productivity. Can we identify who’s at risk and why?

Business goal:

- Give HR insights to proactively retain talent.

Metric of interest:

- Focusing on recall for minority class (attrition = Yes)

Data & Pre-Processing

Dataset:

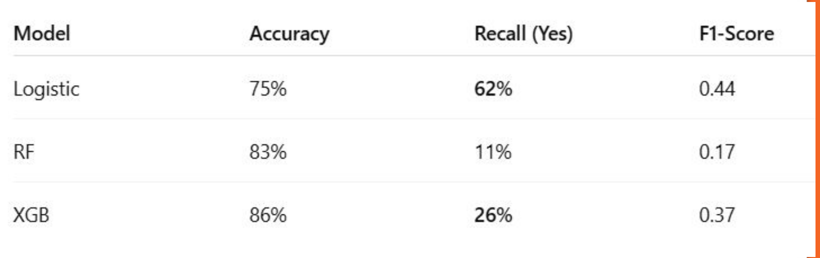
- IBM HR dataset: 1,470 records, 35 features
- Pulled from: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Preprocessing steps:

- Dropped ID and irrelevant columns
- One-hot encoding
- Train/test split with stratification

Modeling Approach

- Logistic Regression (class-weighted)
  - Used class_weight='balanced' to handle class imbalance
  - Best at identifying attrition cases (highest recall)
  - Coefficients provided interpretable feature importance
- Random Forest Classifier
  - Tree-based ensemble model with default settings
  - High overall accuracy, but struggled with recall on attrition cases
  - Good for capturing non-linear relationships, but underperformed on minority class
- XGBoost Classifier
  - Gradient-boosted trees optimized for performance
  - Best overall accuracy and balanced metrics
  - Captured complex patterns while improving recall over Random Forest

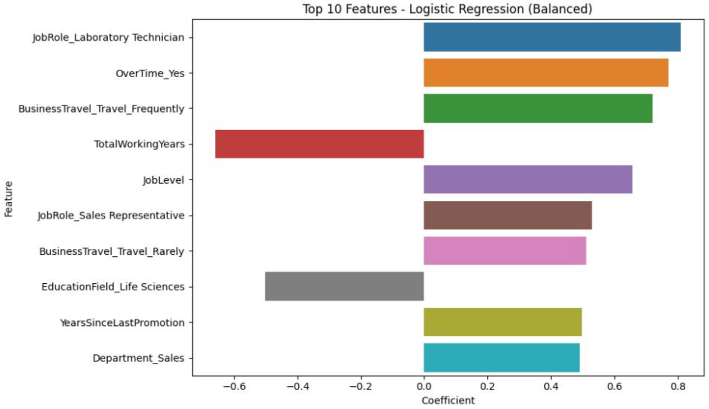


Model	Accuracy	Recall (Yes)	F1-Score
Logistic	75%	**62%**	0.44
RF	83%	11%	0.17
XGB	86%	**26%**	0.37

Key Insights

📌 Feature Importance & Findings

- Top drivers of attrition: Overtime, Job Role, Distance from Home
- Logistic Regression had highest recall
- XGBoost balanced overall accuracy with minority detection



Recommendations

- Improve work-life balance (reduce mandatory overtime)
- Target interventions for high-risk roles
- Consider commute support (flexible schedules, relocation assistance)

Focus*: Reduce false negatives—catch those likely to quit before they do.*