Predicting Employee Attrition Using Machine Learning

1. Stakeholder

The stakeholder for this project is the Human Resources (HR) department of a mid-to-large-sized organiza
They are concerned about employee turnover and want to predict which employees are at risk of leaving.

2. Problem Statement

Employee attrition is costly for businesses due to recruitment, training, and lost productivity.
The HR department wants a predictive model that identifies employees at risk of leaving, allowing proactive

3. Dataset

The dataset used for this project is the "IBM HR Analytics Employee Attrition & Performance" dataset, publ
The dataset can be accessed at: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-employ
The dataset contains 35 attributes, including employee demographics, job roles, and satisfaction metrics.

4. Models Used and Rationale

We experimented with two different types of models:

- Random Forest Classifier: Chosen for its robustness, ability to handle feature importance analysis, and capacity to deal with both categorical and numerical data.
- Logistic Regression: Chosen as a baseline model due to its interpretability and efficiency in handling bi

Each model was tuned using three different hyperparameter settings:

- Random Forest: Number of trees (100, 200, 500)
- Logistic Regression: Regularization parameter C (0.01, 1, 10)

5. Feature Selection & Engineering

Feature selection was based on exploratory data analysis and domain knowledge. Selected features inclu

- Existing Features: Age, JobSatisfaction, YearsAtCompany, WorkLifeBalance, BusinessTravel, and Tota
- Engineered Features:
  - JobTenure = YearsAtCompany / (TotalWorkingYears + 1) to normalize tenure.
  - Satisfaction Score = (JobSatisfaction + WorkLifeBalance) / 2 to create an overall satisfaction metric.

6. Model Evaluation

The models were evaluated using:

- Accuracy: Measures overall correctness.
- Precision: Important to avoid false positives in predicting attrition.
- Recall: Ensures employees likely to leave are identified.
- F1-Score: Balances precision and recall.

7. Future Work

- Incorporating additional external factors such as market conditions or employee feedback surveys.
- Testing advanced models like Gradient Boosting or Neural Networks.
- Feature selection refinement using SHAP values for better interpretability.

8. Recommendation

Based on our analysis, the Random Forest model with optimized hyperparameters performed best in balar
It is recommended for deployment with further validation in real-world HR scenarios.

A link to the complete code and analysis can be found at: https://github.com/your-repo-link.