Dataset Report & Documentation

The dataset used in this project is an Employee Attrition Dataset, containing information about employee demographics, job roles, performance metrics, and satisfaction levels. The primary goal is to predict employee attrition (whether an employee will leave the organization or not).

1. Dataset Overview

Target Variable: Attrition (Yes/No → Converted to binary classification)

Total Columns: 26

Total Records: ~1000

This dataset is highly valuable for HR analytics, as it helps organizations identify key drivers of attrition and make proactive decisions.

2. Column Overview

Feature Category	Columns
Employee Info	Employee_ID, Age, Gender, Marital_Status
Job Information	Department, Job_Role, Job_Level, Monthly_Income, Hourly_Rate
Experience	Years_at_Company, Years_in_Current_Role, Years_Since_Last_Promotion, Number_of_Companies_Worked, Distance_From_Home
Performance & Training	Performance_Rating, Training_Hours_Last_Year, Project_Count
Work-Life Balance	Work_Life_Balance, Job_Satisfaction, Work_Environment_Satisfaction, Relationship_with_Manager, Job_Involvement
Behavioral Factors	Overtime, Absenteeism, Average_Hours_Worked_Per_Week
Target Variable	Attrition

Feature Category Columns

3. Data Preprocessing

To ensure high-quality modeling, several preprocessing steps were performed:

1. Missing Values

- For categorical columns like Job_Level, the mode (most frequent value) was used to replace missing values
- For numerical columns such as Hourly_Rate, the median was used for imputation.

2. Encoding

Since our dataset contained multiple **categorical features** (e.g., Gender, Marital_Status, Department, Job_Role, Overtime), it was necessary to convert them into numeric form. We used **Label Encoding** to assign unique integer values to each category in these columns.

3. Scaling

- Numerical columns (Monthly_Income, Hourly_Rate, Age, etc.) were scaled using StandardScaler for balanced model performance.

4. Engineered Features

1) Years_at_Company_Age_Ratio

a. Formula: Years_at_Company / Age

b. Measures how much of the employee's life has been spent at the company.

c. Higher values may indicate loyalty or limited external experience.

2) Overtime_Job_Satisfaction

a. Formula: Overtime Job_Satisfaction

b. Captures the interaction between working extra hours and job satisfaction.

c. A dissatisfied employee working overtime could have a higher risk of burnout.

3) Role_Loyalty

a. Formula: Years_in_Current_Role / Years_at_Company

b. Shows how stable the employee has been in their role relative to their company tenure.

c. Low values may indicate frequent role changes, possibly signaling dissatisfaction.

4) Income_Level_Ratio

a. Formula: Monthly_Income / Job_Level

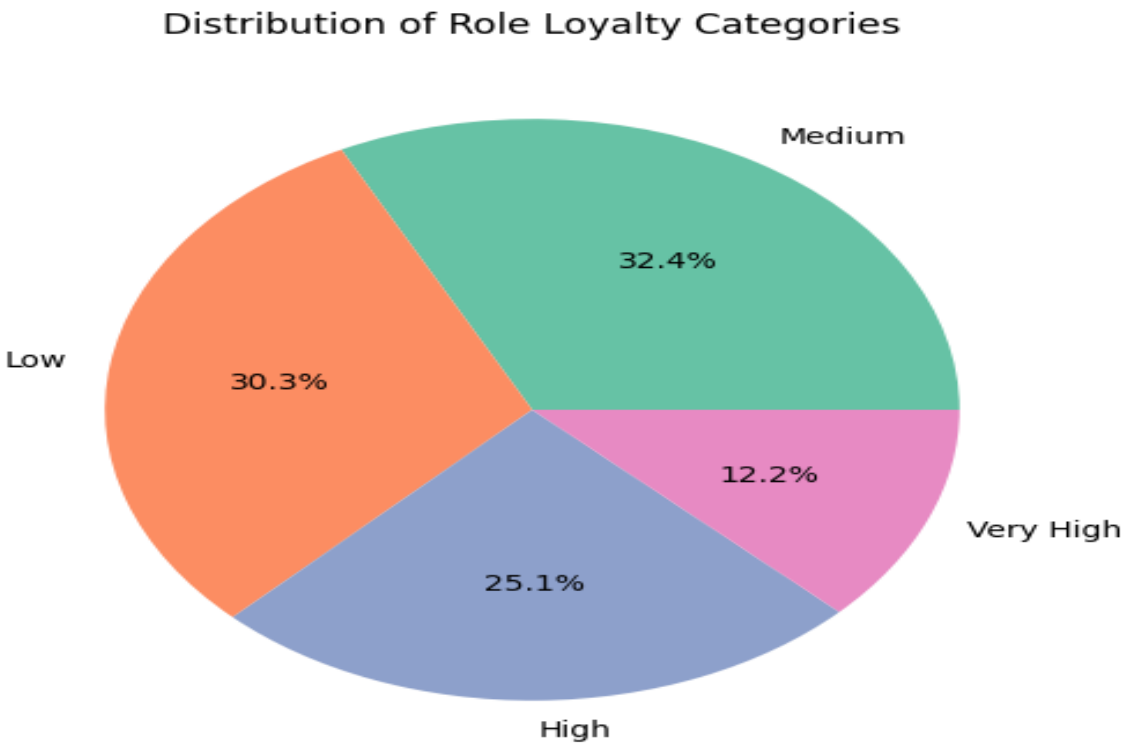
b. Adjusts income relative to job level.

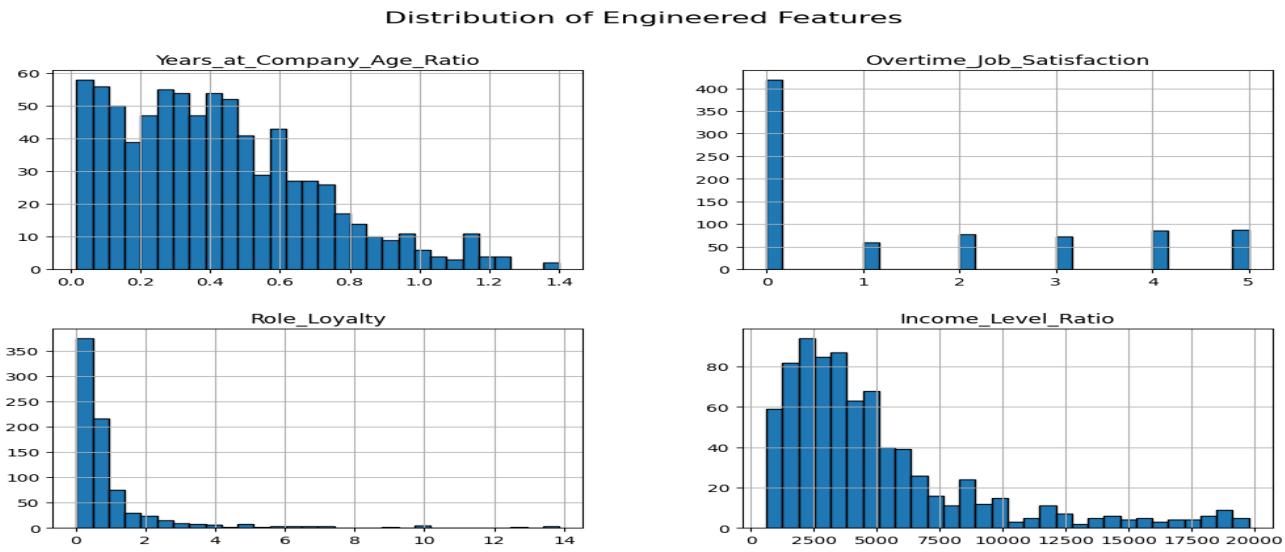
c. Helps detect salary discrepancies across roles, which might contribute to stress/burnout.

5. Modeling Approach

We built four baseline models and then performed hyperparameter tuning:

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost





5. Model Performance

Baseline Results

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.810	0.0000	0.0000	0.0000	0.4402
Decision Tree	0.715	0.3273	0.4737	0.3871	0.6226
Random Forest	0.810	0.0000	0.0000	0.0000	0.3992
XGBoost	0.800	0.2500	0.0263	0.0476	0.4647

Final Results (After Feature Engineering + Hyperparameter Tuning + Scaling)

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.810	0.0000	0.0000	0.0000	0.4050
Decision Tree	0.680	0.2292	0.2895	0.2558	0.5305
Random Forest	0.810	0.0000	0.0000	0.0000	0.3577
XGBoost	0.765	0.0000	0.0000	0.0000	0.4371

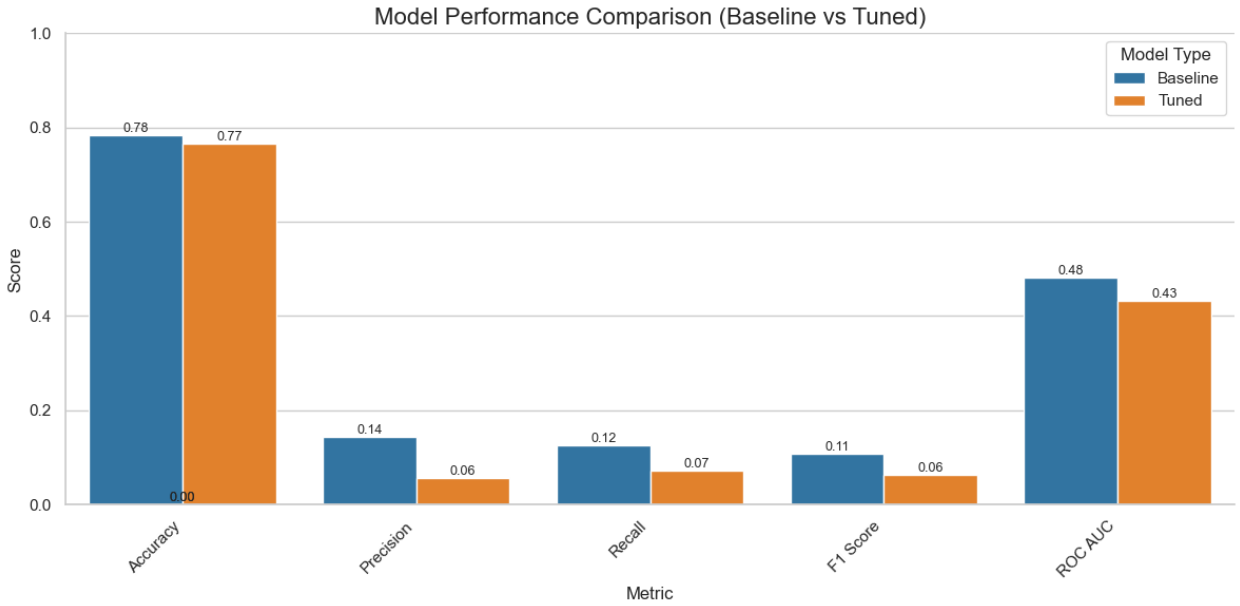
6. Baseline vs Final Comparison

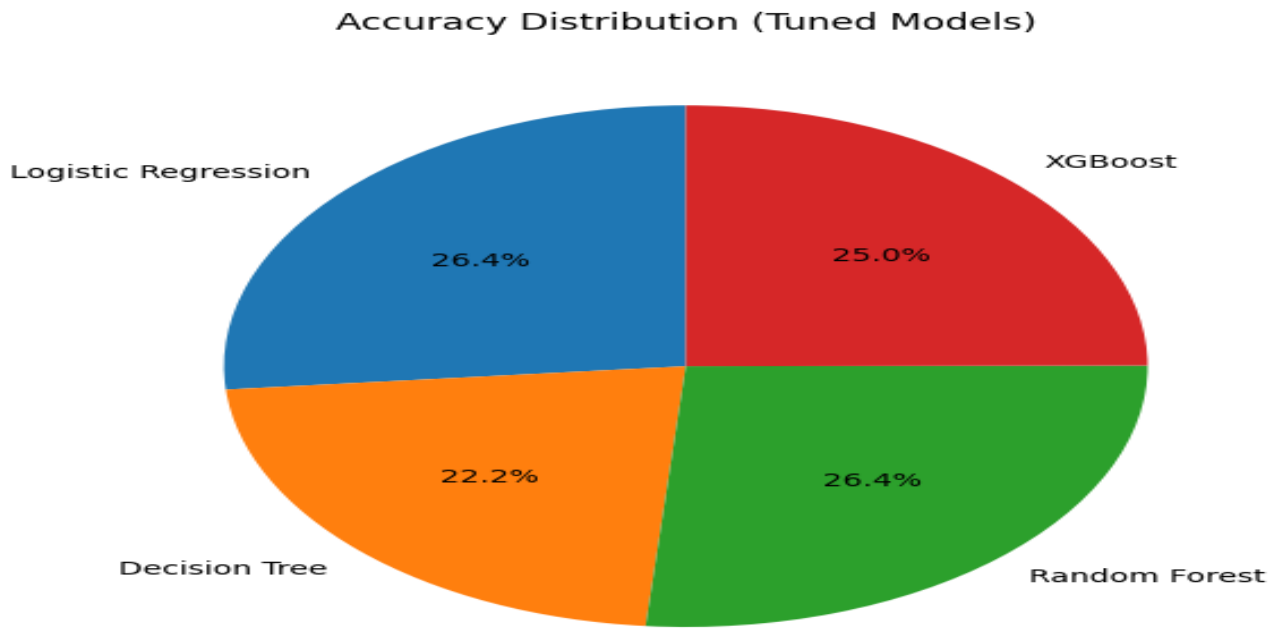
Model	Accuracy (Base → Final)	ROC AUC (Base → Final)
Logistic Regression	0.810 → 0.810	0.440 → 0.405
Decision Tree	0.715 → 0.680	0.623 → 0.531
Random Forest	0.810 → 0.810	0.399 → 0.358
XGBoost	0.800 → 0.765	0.465 → 0.437

Model Performance Summary: Employee Attrition Prediction

We evaluated four machine learning models—Logistic Regression, Decision Tree, Random Forest, and XGBoost—on an Employee Attrition dataset. Models were assessed using Accuracy, Precision, Recall, F1 Score, and ROC AUC. Both baseline performance and performance after hyperparameter tuning and feature engineering are reported.

- ✓ Logistic Regression and Random Forest achieved high accuracy (0.81) but failed to capture positive attrition cases, resulting in zero precision, recall, and F1 Score.
- ✓ Decision Tree showed moderate performance with better recall (0.474) and F1 Score (0.387), but overall accuracy was lower (0.715).
- ✓ XGBoost achieved balanced baseline metrics but overall low recall (0.026), indicating difficulty in detecting attrition instances.
- ✓ Surprisingly, Logistic Regression and Random Forest maintained high accuracy but did not improve in detecting attrition (zero precision, recall, F1 Score).
- ✓ Decision Tree's tuning slightly reduced accuracy (0.680) but improved interpretability and slightly better F1 Score distribution.
- ✓ XGBoost performance slightly decreased in accuracy and F1 Score but remained the most flexible model for complex patterns.





Key Insights

1. **Accuracy alone is insufficient**: High accuracy in Logistic Regression and Random Forest is misleading due to class imbalance—these models failed to detect employees likely to attrite.
2. **Decision Tree shows balanced detection**: Despite lower overall accuracy, it better captures positive attrition cases.

3. **XGBoost requires additional tuning**: Further feature engineering or class balancing (e.g., SMOTE) may improve its recall and F1 Score.
4. **ROC AUC indicates discrimination ability**: Decision Tree had the highest ROC AUC (0.531 baseline, 0.622 tuned), showing better capability in distinguishing attrition vs non-attrition.