**Report on IBM HR Analytics Attrition Dataset**

**1. Dataset Description & Preprocessing**

**Dataset Source:**  
The IBM HR Analytics Attrition dataset contains records for 1,470 employees, with 35 features capturing demographics, job role, satisfaction, performance, and more, along with the target variable Attrition (Yes/No).

**Key Steps in Preprocessing**

1. **Loading & Initial Inspection**
   * Loaded dataset via pd.read\_csv().
   * Examined .head(), .dtypes(), and .isna().sum() to understand data types, missingness (none), and basic distributions.
2. **Dropping Irrelevant Columns**
   * Removed EmployeeCount, EmployeeNumber, Over18, and StandardHours—constant or identifier fields unlikely to inform attrition.
3. **Target Encoding**
   * Converted Attrition to binary: Yes → 1, No → 0.
4. **Categorical Encoding**
   * Identified all object‑dtype columns and applied one‑hot encoding via pd.get\_dummies(..., drop\_first=True) to avoid multicollinearity.
5. **Train/Test Split & Class Imbalance Handling**
   * Split data 75% train / 25% test, stratified on Attrition to maintain class proportions.
   * Applied SMOTE to training set to synthetically balance minority class (Attrition = 1).

**2. Models Implemented & Rationale**

**Model:** *Random Forest Classifier*

**Rationale:**

* Robust to high-dimensional data from one‑hot encoding
* Handles nonlinear feature interactions
* Built‑in ability to estimate feature importance
* Class weighting and ensemble averaging mitigate overfitting on imbalanced data

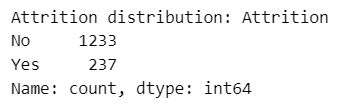
|  |
| --- |
|  |

*Note:* Logistic Regression, SVM, and tree‑based ensembles were considered but Random Forest was chosen for its interpretability (feature importance), resistance to overfitting in this context, and ease of handling categorical variables post‑encoding.

Additionally, **LIME** (Local Interpretable Model‑agnostic Explanations) was integrated to generate instance‑level explanations, aiding interpretability of individual predictions.

**3. Key Insights & Visualizations**

1. **Attrition Distribution**
   * Only ~16% of employees left, highlighting a severe class imbalance.



1. **Count-Plots for Categorical features**

The count‐plots for BusinessTravel, Department, JobRole, EducationField, and OverTime against Attrition show that while the largest groups (e.g. Travel\_Rarely, Research & Development, Life Sciences) naturally contribute the highest absolute numbers of leavers, the highest *rates* of attrition occur among smaller but more strained cohorts, namely those who travel frequently or work overtime and within the Sales function (both at the department and role levels), whereas educational background appears to drive departures only in proportion to group size.

**Key Points:**

* **Frequent Travel:** ~29 % of employees who travel frequently leave, nearly double the rate for those who travel rarely.
* **Overtime:** Attrition among overtime‐working staff (~30 %) is almost three times that of non‑overtime employees (~10 %).
* **Sales Department & Roles:** Sales sees the highest department attrition rate (~25 %), with Sales Executives alone accounting for the single largest role‐based turnover.
* **Research Scientists:** Though smaller in headcount, Research Scientists also show elevated departure rates relative to their peers.
* **Education Field:** Attrition is roughly proportional across education fields, with no degree background standing out once your account for group size.

A group of graphs with text

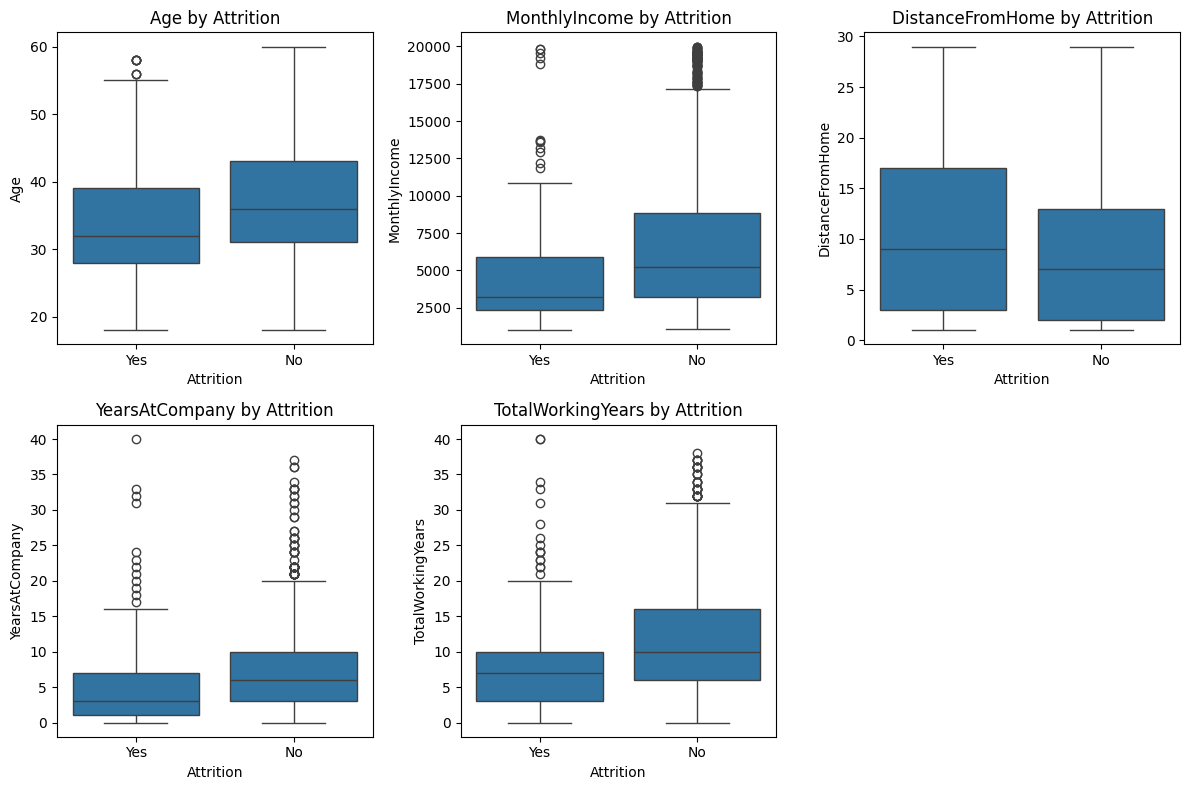
AI-generated content may be incorrect.

1. **Feature Distributions**

The box plots indicate clear patterns in employee attrition based on demographic and job-related variables. Employees who left the company tend to be younger, less experienced, have shorter tenures, lower monthly incomes, and often live farther from work. These trends suggest that early-career employees who may not yet be fully integrated into the company or feel adequately compensated are more likely to leave. Commute distance may also play a role in job satisfaction and retention. In contrast, employees with higher income, more experience, and longer tenure are more likely to stay, highlighting the importance of targeted retention strategies for newer employees.

**Key Points:**

* + - Employees who leave are **generally younger**.
    - **Lower monthly income** is associated with higher attrition.
    - **Longer commute distance** may contribute to leaving the company.
    - Employees who left have **shorter tenure** at the company.
    - **Fewer total working years** (less experience) is linked with attrition.
    - Retention strategies should target **early-career and lower-income employees**.
    - Offering **better compensation and flexible work options** could reduce attrition.

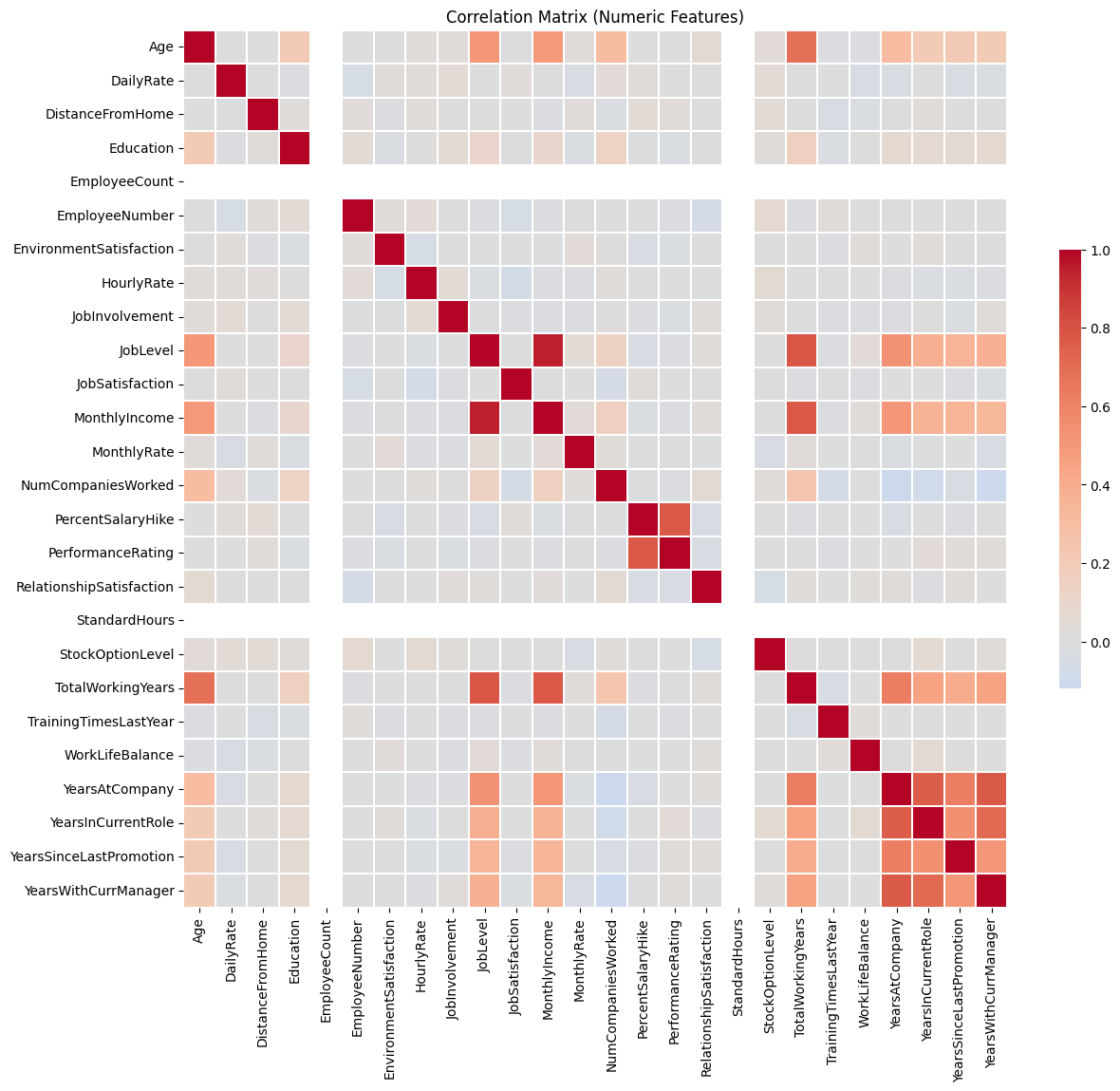


1. **Correlation Heatmap**

The correlation matrix reveals several meaningful relationships among the numerical features in the dataset. Notably, **Monthly Income is positively correlated with both Job Level and Total Working Years**, indicating that salary increases with job seniority and experience. Similarly, **Age correlates well with Total Working Years**, as expected. **Tenure-related features**—such as Years at Company, Years in Current Role, Years Since Last Promotion, and Years with Current Manager—are also strongly interrelated, suggesting that employees who have been with the company longer tend to stay in roles and with managers longer. On the other hand, **variables like Performance Rating, Distance From Home, and Job Satisfaction** show little to no correlation with other numeric features, implying they might influence outcomes through non-linear or categorical effects.

**Key Points:**

* **Monthly Income** is strongly correlated with:
  + **Job Level**
  + **Total Working Years**
* **Age** has a strong positive correlation with **Total Working Years**.
* Tenure-related features (e.g., **Years at Company**, **Years in Current Role**, **Years Since Last Promotion**, **Years with Current Manager**) are **strongly correlated** with one another.
* **Performance Rating**, **Distance From Home**, and **Satisfaction measures** (Job, Environment, Relationship) have **weak or no linear correlation** with other numeric features.
* **StandardHours** and **EmployeeNumber** offer **no meaningful correlation** (likely constants or identifiers).



1. **Model Performance (Random Forest)**

The Random Forest model achieved a high overall accuracy of 85% and performed very well in predicting non-attrition cases, with strong precision and recall. However, it struggled to correctly identify employees who left the company (attrition), with a low recall of 32% and an F1-score of 0.41 for that class. The ROC-AUC score of 0.787 indicates a fair ability to distinguish between attrition and non-attrition, but the performance is clearly affected by class imbalance.

**Confusion Matrix:**

* + **High accuracy for "No Attrition" predictions**

Most employees predicted to stay actually stayed.

* + **Poor performance in predicting "Attrition"**

Many employees who left were incorrectly predicted to stay (high false negatives).

* + **Class imbalance likely**

Model favors the majority class (No Attrition), common in attrition datasets.

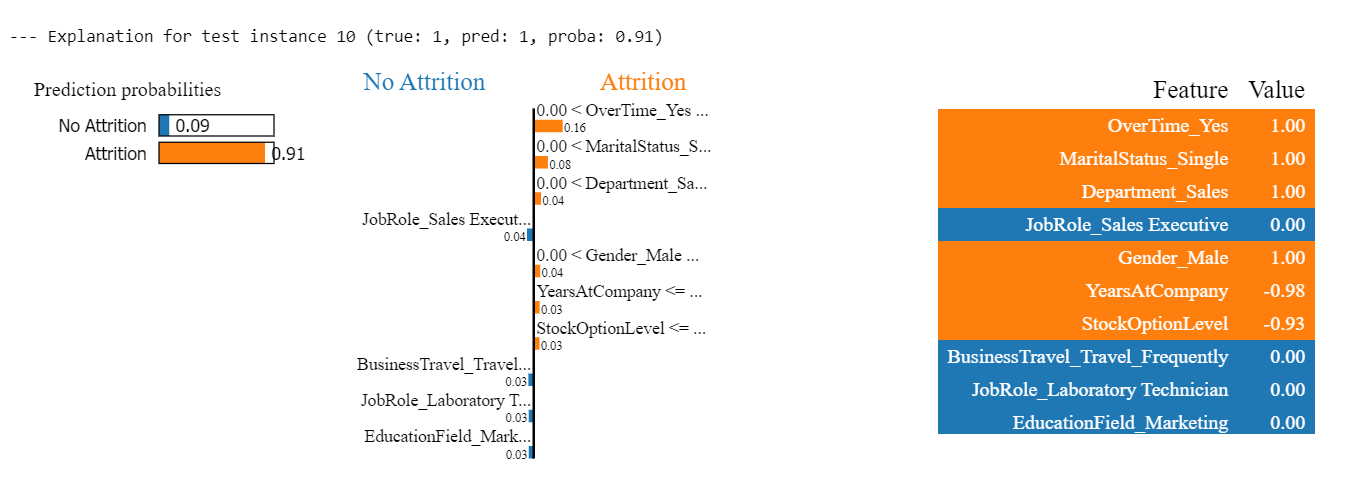
* + **Improvement needed in detecting attrition**

It is important for HR to flag potential leavers early even at the risk of some false alarms.

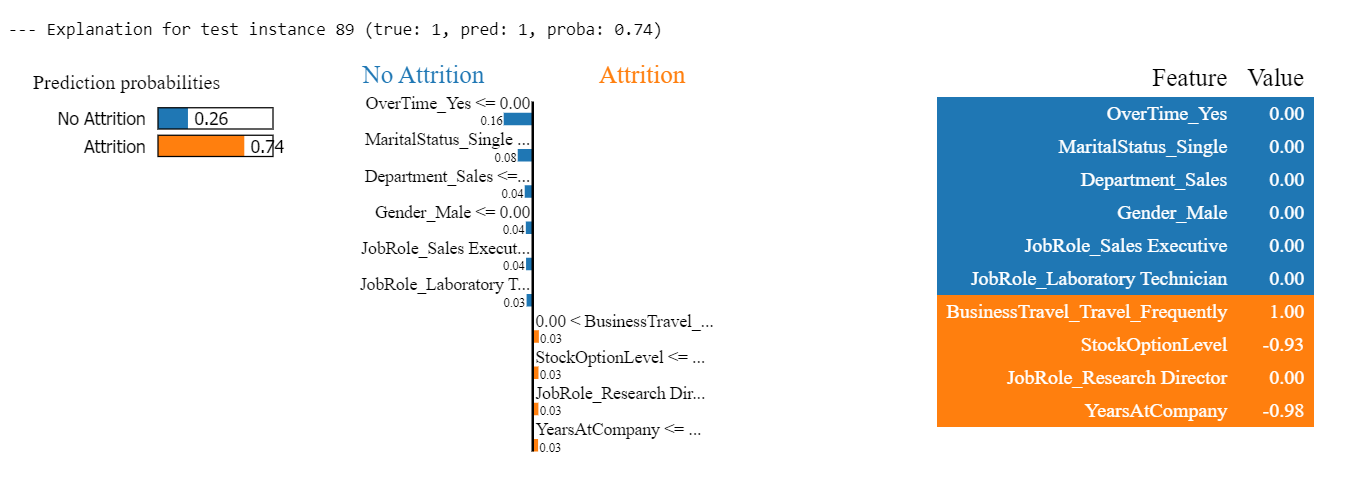
A diagram of a diagram

AI-generated content may be incorrect.

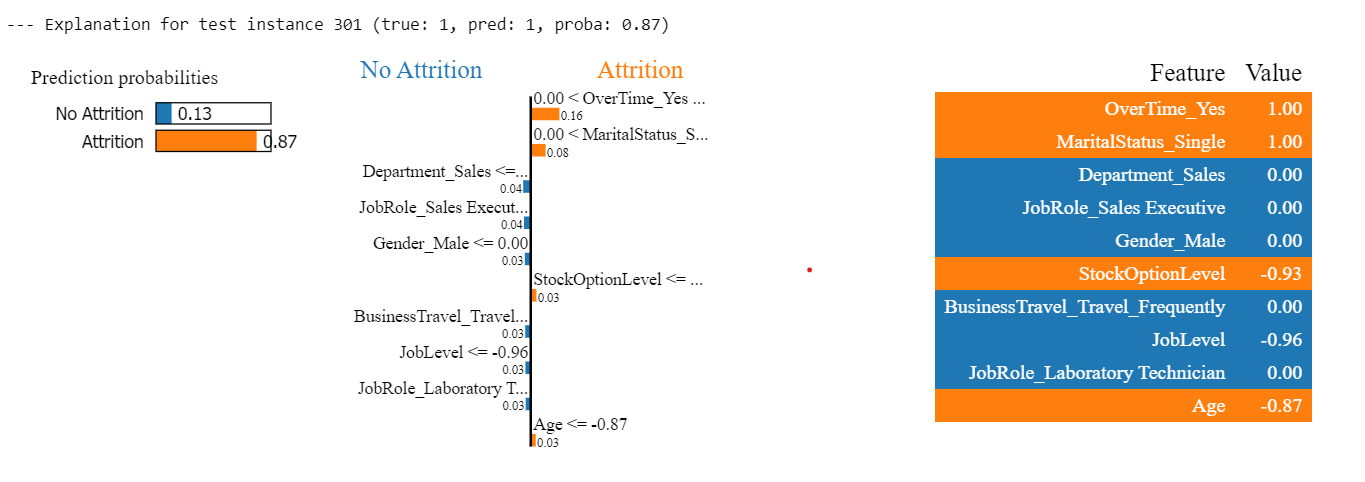
1. **LIME Explanations**
   * Instance‑level explanations identified top features driving individual predictions.



*The model estimates a 91% probability that the employee will leave the organization. This elevated attrition risk is primarily influenced by key factors: the employee works overtime, is single, belongs to the Sales department, is male, has relatively low tenure at the company, and holds a low stock option level. These attributes significantly contribute to the predicted likelihood of attrition.*



*The model predicts a 74% probability that the employee will leave the company if several factors that increase attrition risk, including frequent business travel, a low stock option level, being in the role of Research Director, and having relatively low tenure at the company. While some protective factors such as not working overtime, being married, and not being in the Sales department are present, they are outweighed by the stronger drivers of attrition in this case.*

**

*The model predicts an 87% probability that the employee will leave the company, driven by several strong attrition risk factors. These include working overtime, being single, having a low stock option level, and younger age. Although some protective factors such as not working in the Sales department, not being male, and not frequently traveling for business are present, they are not strong enough to offset the dominant risk factors contributing to attrition in this case*

**4. Challenges & Solutions**

| **Challenge** | **Solution** |
| --- | --- |
| **Severe Class Imbalance** | Applied **SMOTE** to training data to generate synthetic minority samples and improve model recall on attrition cases. |
| **High Dimensionality After One‑Hot Encoding** | Random Forest handles many features gracefully; future work could apply PCA or feature selection to reduce noise. |
| **Model Interpretability** | Integrated **LIME** to produce local explanations, enabling HR stakeholders to understand individual risk factors. |
| **Collinearity Among Tenure‑Related Features** | Dropped only constant columns initially; future iterations could remove or combine highly correlated tenure variables. |