**Employee Attrition Prediction**

**Introduction**

Employee attrition is a critical concern for many organizations, as losing skilled employees can impact productivity, morale, and costs. In this project, we leverage data analytics and machine learning to understand patterns behind employee attrition using the **IBM HR Analytics Dataset**. We aim to:

* Explore and preprocess the dataset
* Build an interpretable classification model
* Extract actionable insights
* Visualize key patterns

1. **Dataset Description and Preprocessing**

The dataset contains information about **1,470 employees** across various attributes such as job role, salary, experience, and work-life balance. The target variable is **"Attrition"**, which indicates whether an employee has left the company.

**Target Variable**: Attrition (Yes / No)

**Download Dataset:** [**https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset**](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset)

**Key Features**

* **Demographic**: Age, Gender, MaritalStatus
* **Work Experience**: TotalWorkingYears, YearsAtCompany, YearsInCurrentRole
* **Job Details**: JobRole, Department, BusinessTravel
* **Compensation**: MonthlyIncome, StockOptionLevel
* **Satisfaction**: JobSatisfaction, WorkLifeBalance
* **Others**: DistanceFromHome, OverTime, EnvironmentSatisfaction

**Preprocessing Steps**

1. **Column Removal**: Dropped non-informative columns like EmployeeCount, EmployeeNumber, Over18, and StandardHours.
2. **Label Encoding**: Converted binary categorical columns such as Attrition, Gender, and OverTime to numerical values (0 and 1).
3. **One-Hot Encoding**: Applied get\_dummies on multi-class categorical columns like JobRole, MaritalStatus, Department, etc.
4. **Feature Scaling**: Standardized numerical features using StandardScaler to ensure uniform scale.
5. **Model Implementation and Justification**

### Model Used: Logistic Regression

#### Why Logistic Regression?

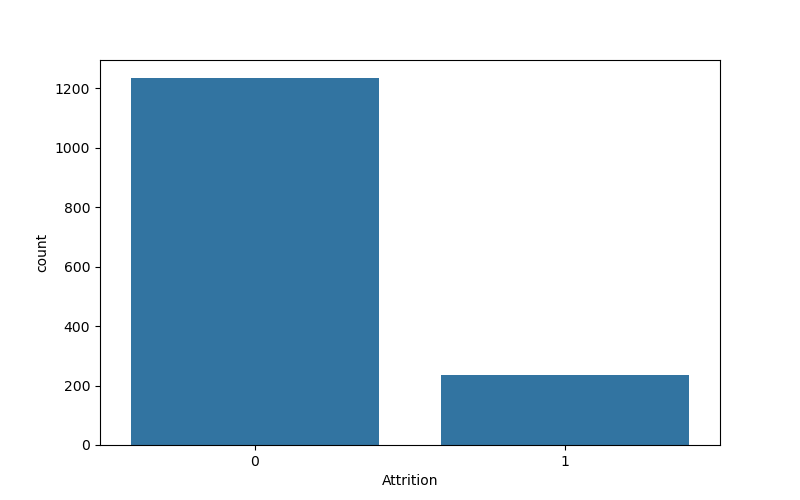
* Suitable for **binary classification** problems.
* Provides **interpretability** via feature coefficients.
* Efficient and quick to train on tabular data.
* Provides **probabilistic outputs** useful for confidence scoring.

### Evaluation Metrics

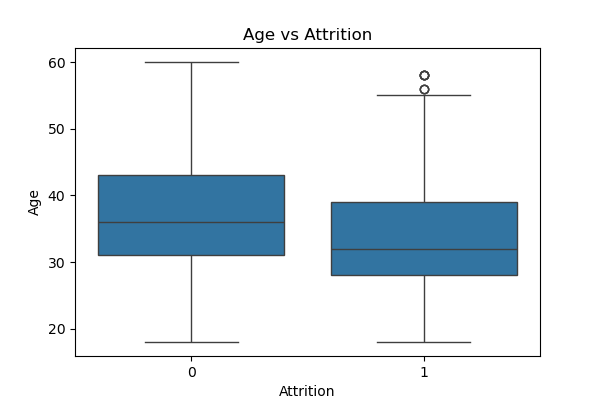
* **Accuracy**: Measures overall correctness.
* **Precision & Recall**: Handle class imbalance.
* **F1-score**: Balance between precision and recall.
* **Confusion Matrix**: Visualizes true vs predicted labels.

1. **Key Visualizations & Insights**

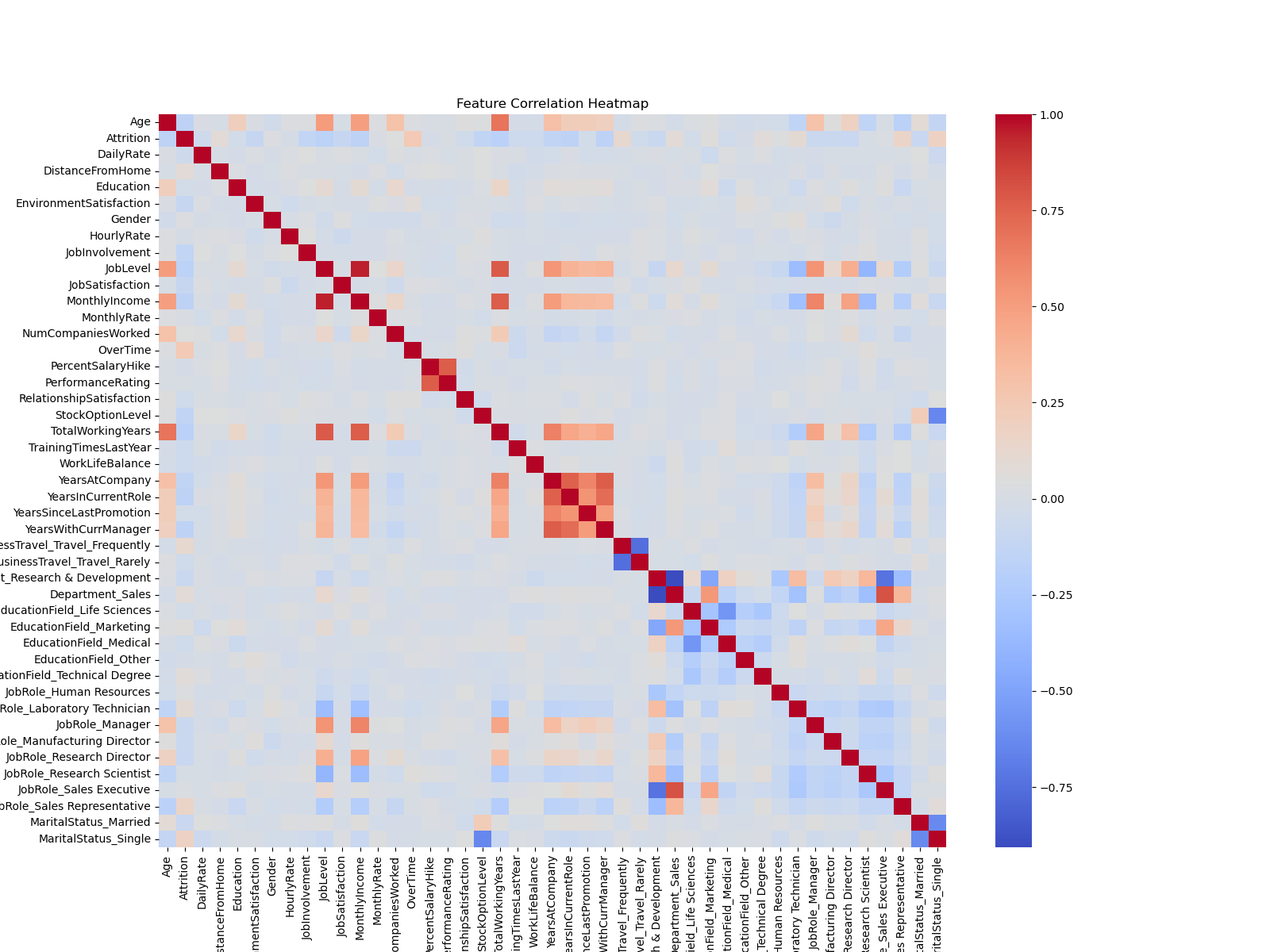
### Attrition Distribution

* Imbalanced dataset: Only ~16% employees left.
* Imbalance handled using stratify during splitting

### Age vs Attrition

* Younger employees show higher attrition rates.
* Suggests need for stronger engagement strategies for newer recruits.

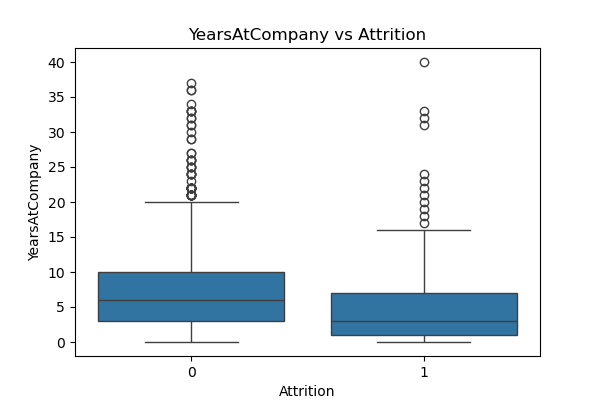
### Correlation Heatmap

* Strong correlation between attrition and features like OverTime, MonthlyIncome, and DistanceFromHome.

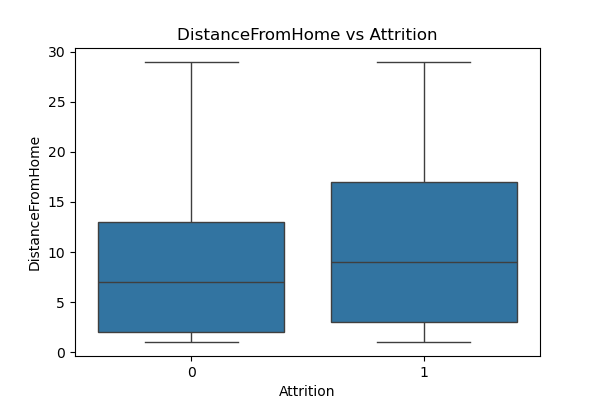
### Total Working Hours vs Attrition

* Employees with fewer working hours tend to leave more.
* Indicates that low engagement or under-utilization might lead to attrition.

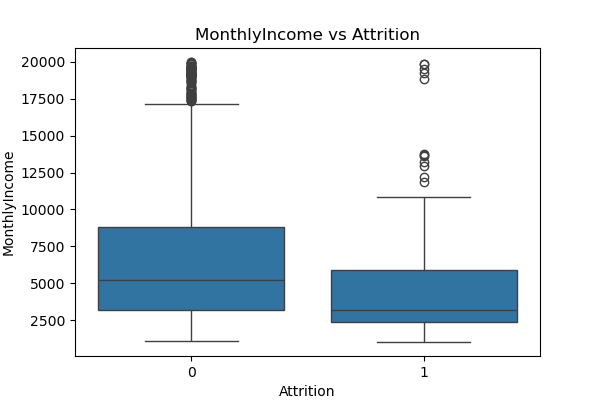
### Years At Company vs Attrition

* Employees with fewer years at the company are more likely to leave.
* Retention programs should target new hires.

### Distance from Home vs Attrition

* Employees living far from the workplace are more prone to attrition.
* Flexible or remote work options can reduce this.

### Monthly Income vs Attrition

* Employees with lower income are more likely to leave.
* Indicates need for salary satisfaction reviews.

1. **Model Explanation with LIME**

We used **LIME (Local Interpretable Model-Agnostic Explanations)** to explain individual predictions:

explainer = lime.lime\_tabular.LimeTabularExplainer(

training\_data=X\_train,

feature\_names=X.columns,

class\_names=['No Attrition', 'Attrition'],

mode='classification'

)

* **Transparent decision-making** helps HR understand which features led to an attrition prediction.
* **Example**: LIME output for one employee might show that long DistanceFromHome, OverTime=Yes, and low JobSatisfaction contributed to predicted attrition.

1. **Challenges Faced & Solutions**

|  |  |  |
| --- | --- | --- |
| **Challenge** | **Description** | **Solution** |
| Class Imbalance | Only ~16% employees left the company | Used stratify=y in train\_test\_split to maintain class ratio |
| Redundant Features | Non-informative columns present | Dropped after analyzing variance and domain knowledge |
| High-Dimensional Encoding | One-hot encoding increased features significantly | Used drop\_first=True to avoid dummy trap and reduce dimensionality |
| Explainability | HR needs to trust model decisions | Integrated LIME for local explanation |

1. **Conclusion and Future Scope**

### Summary

* Built a **logistic regression** model to predict employee attrition.
* Preprocessed data using encoding and scaling.
* Performed **visual analysis** to extract actionable HR insights.
* Used **LIME** to explain model decisions.

### Actionable Insights

* Attrition is **higher among younger**, **lower-paid**, and **overworked** employees.
* Features like OverTime, JobSatisfaction, and MonthlyIncome are key predictors.

### Future Enhancements

* Use **ensemble methods** like Random Forest or XGBoost.
* Apply **SMOTE** or **class weights** to handle imbalance.
* Build a **web dashboard** for HR teams.
* Incorporate **textual data** (e.g., feedback) for deeper insights.