PROPOSAL FOR EFFICIENCY IN HUMAN RESOURCES

By: Nguyen Minh Xuan - HUST

Tools: Python, Pandas, Seaborn, Matplotlib, Scikit-learn

Table of contents

Dataset Overview

```
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 44 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   Department                1470 non-null   object
 4   DistanceFromHome          1470 non-null   int64
 5   Education                 1470 non-null   int64
 6   EducationField            1470 non-null   object
 7   EmployeeCount             1470 non-null   int64
 8   EmployeeNumber            1470 non-null   int64
 9   EnvironmentSatisfaction   1470 non-null   int64
 10  Gender                    1470 non-null   object
 11  PerformanceIndex          1470 non-null   int64
 12  JobInvolvement            1470 non-null   int64
 13  JobLevel                  1470 non-null   int64
 14  JobRole                   1470 non-null   object
 15  JobSatisfaction           1470 non-null   int64
 16  MaritalStatus             1470 non-null   object
 17  MonthlyAchievement        1470 non-null   int64
 18  NumCompaniesWorked        1470 non-null   int64
 19  Over18                    1470 non-null   object
 20  OverTime                  1470 non-null   int64
 21  PerformanceRating         1470 non-null   int64
 22  RelationshipSatisfaction  1470 non-null   int64
```

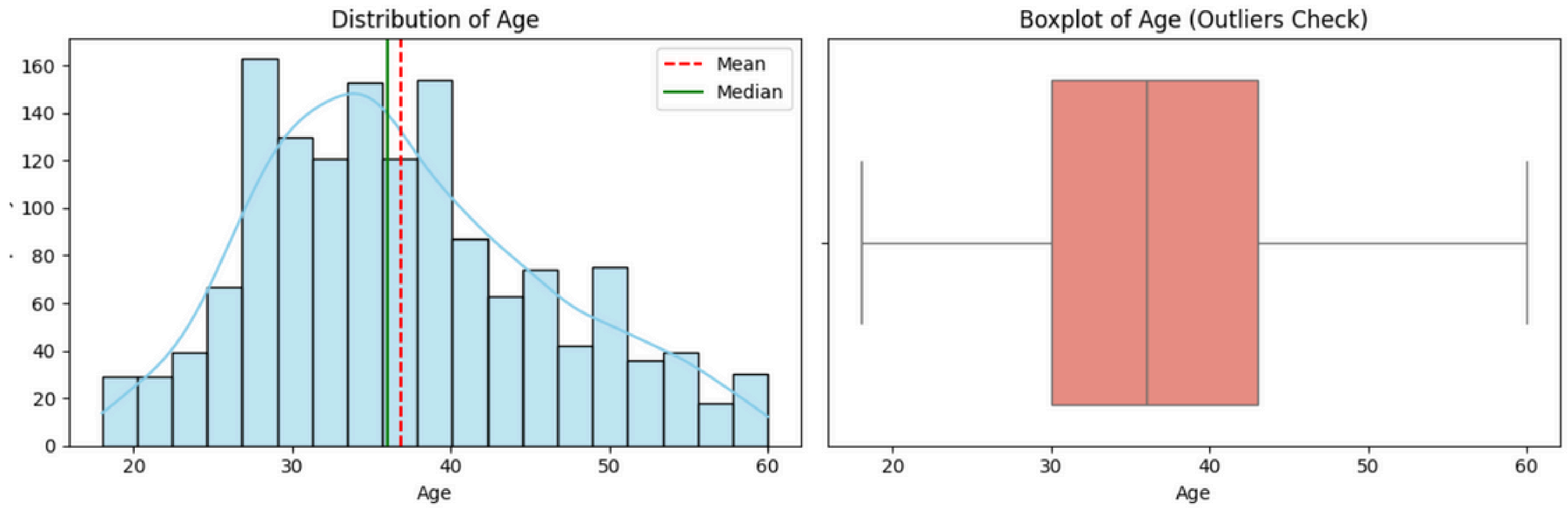
Dataset Description:

- ~1500 records (HR data from 2023 & 2024)
- ~40 employee-related features (demographics, job information, satisfaction levels)
- Target variable: Attrition (Yes/No)
- No missing values or critical duplicates
- Mix of numerical and categorical data

Data Cleaning & Preprocessing

- Removed irrelevant fields (EmployeeCount, StandardHours, Over18)
- Encoded categorical variables (Label / OneHot)
- Standardized numerical fields for Logistic Regression
- Train–test split (80% / 20%)

Univariate Analysis (Numerical Features)

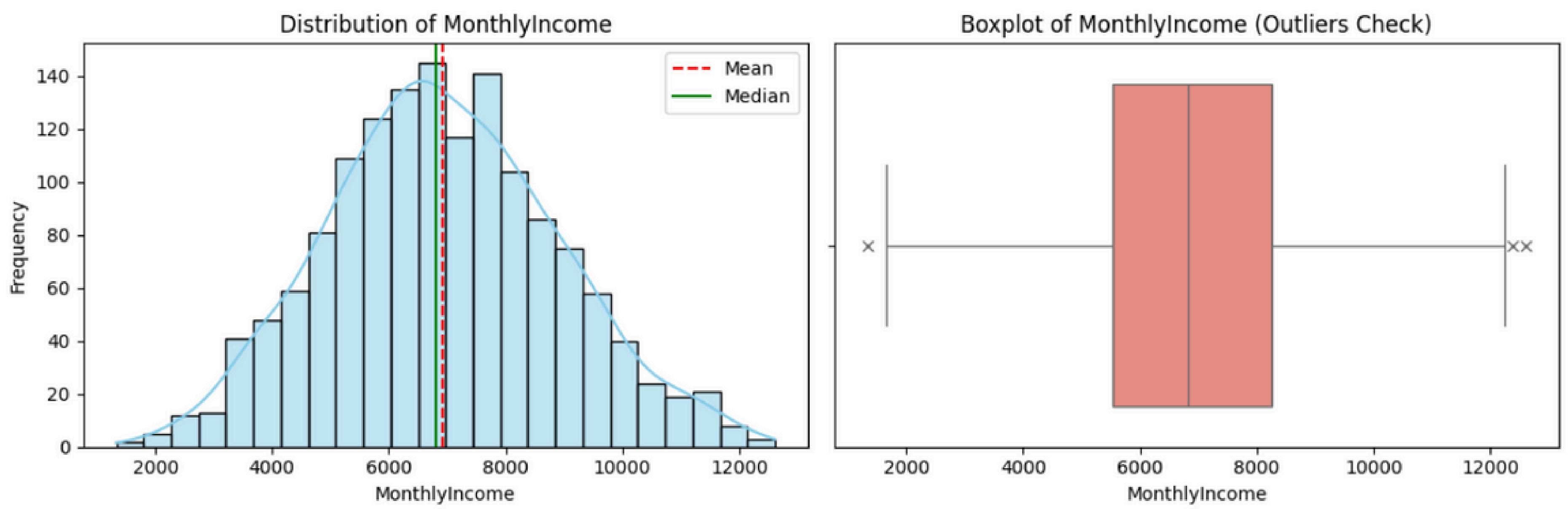


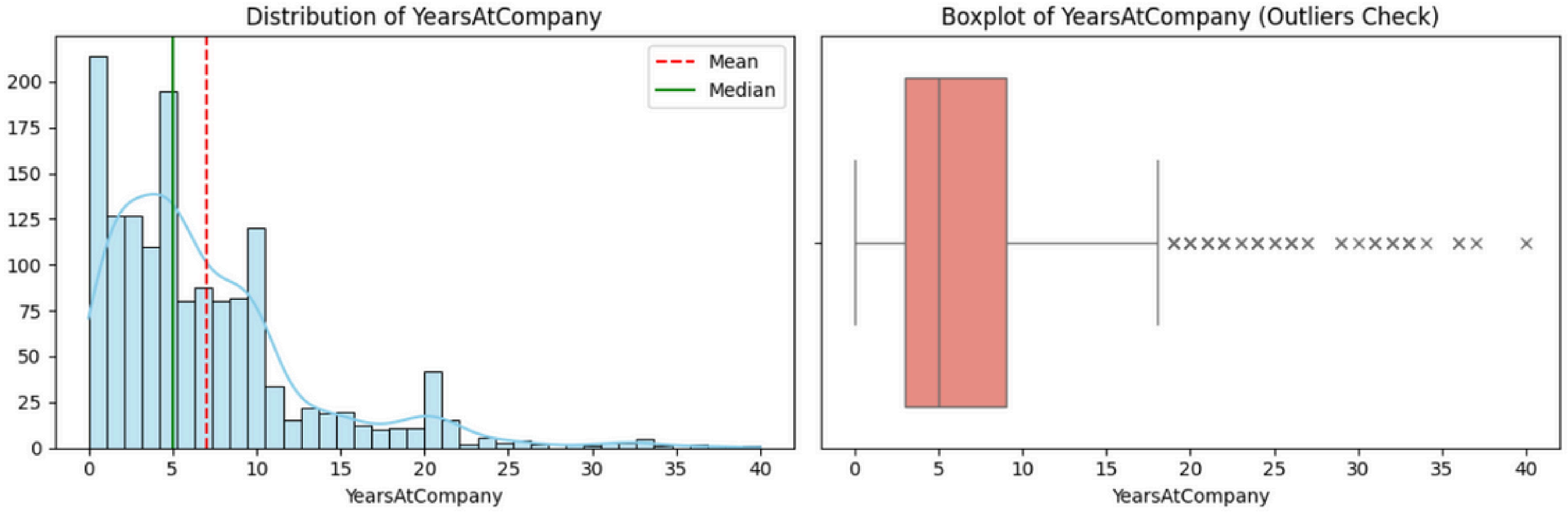
Examples:

- Age
- Monthly Income
- Years at Company

Key Observations:

- Employees who left tend to be younger
- Lower income groups show higher attrition
- Low tenure employees are at highest risk

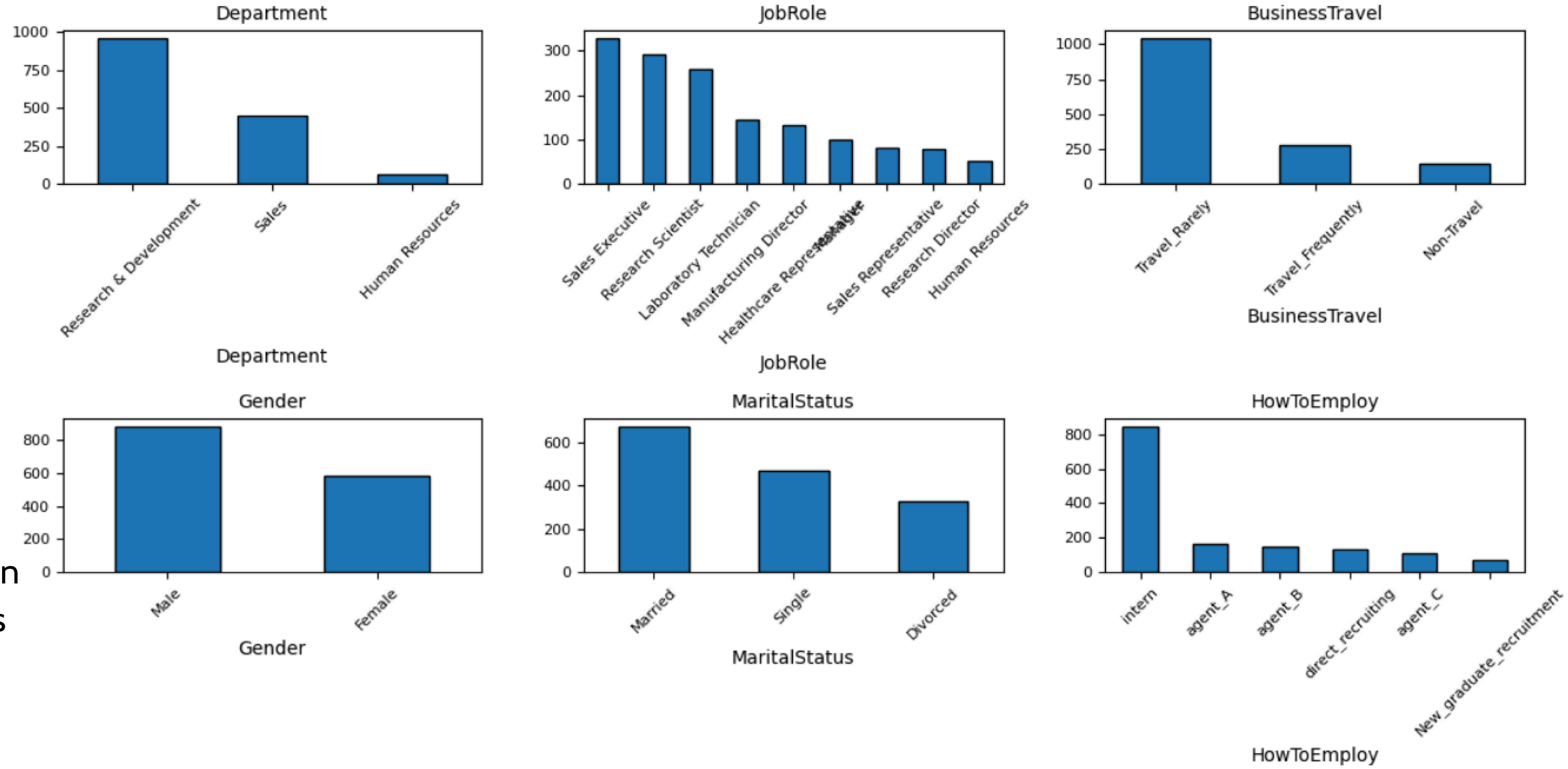




Univariate Analysis (Categorical Features)

Important features explored:

- Department
- Job Role
- Business Travel
- Marital Status
- Employment Type (HowToEmploy)



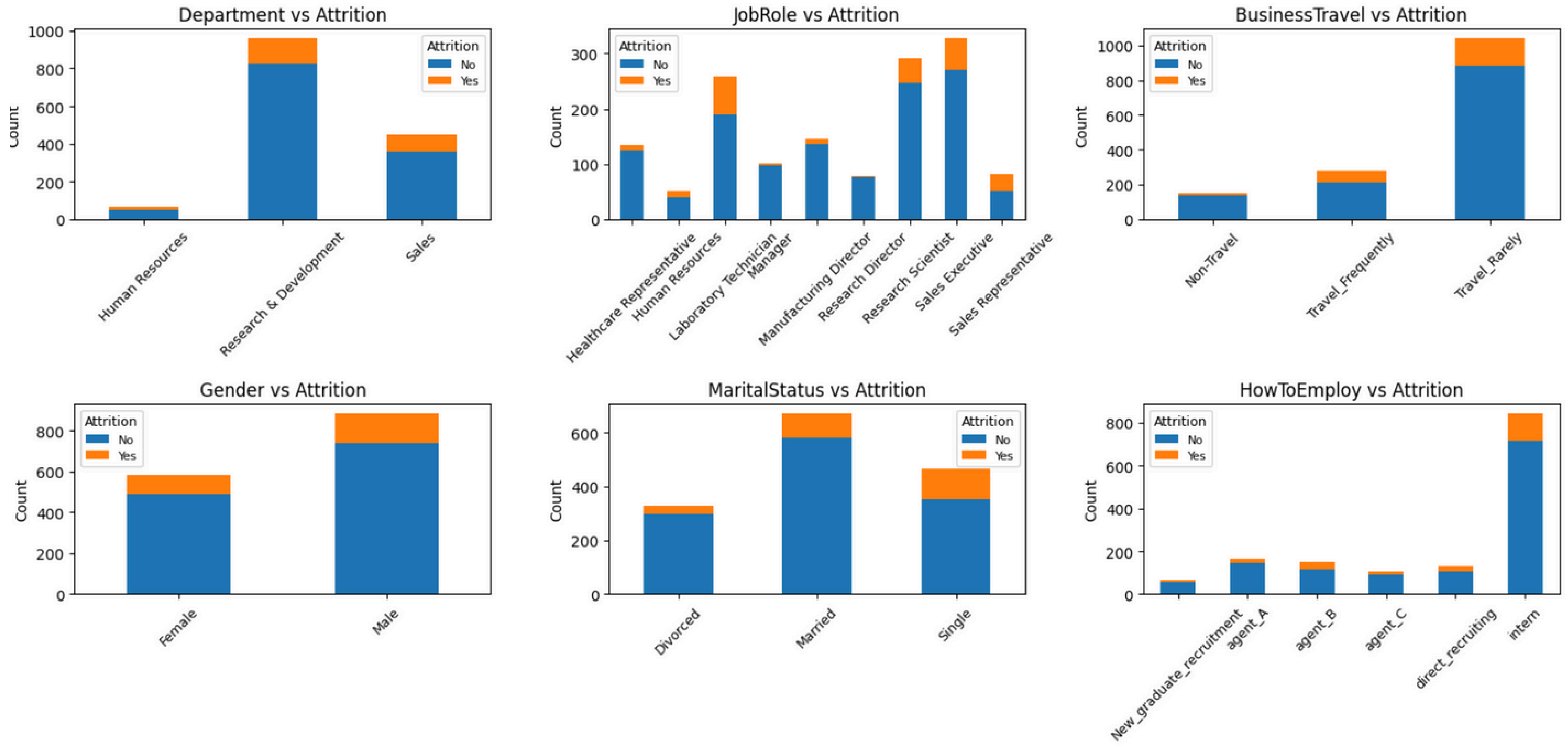
Insight:

- Sales roles show higher attrition
- “Travel Frequently” employees resign more
- Single employees have higher turnover rates

Attrition vs Categorical Features

Observations

- Sales Representatives → highest attrition segment
- Employees hired via recruitment agents → more likely to leave
- Interns show very low attrition, indicating strong loyalty

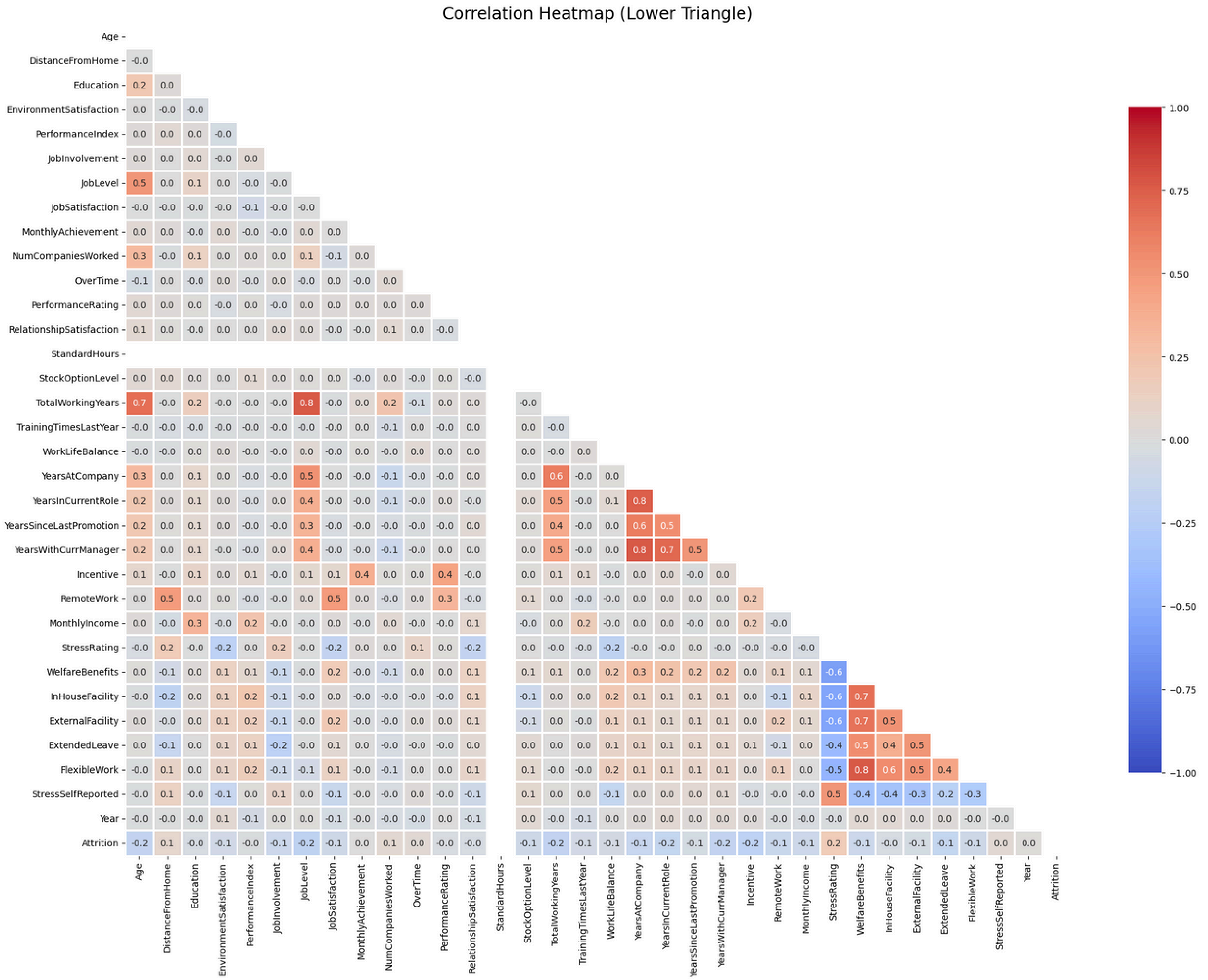


Correlation Heatmap

Takeaways:

- Attrition is weakly correlated with all numerical variables
- Slight negative correlations: Age, Income, JobLevel, Tenure
- Slight positive correlations: StressRating, DistanceFromHome

→ Indicates that attrition is a multi-factor problem, requiring ML models rather than simple thresholds.



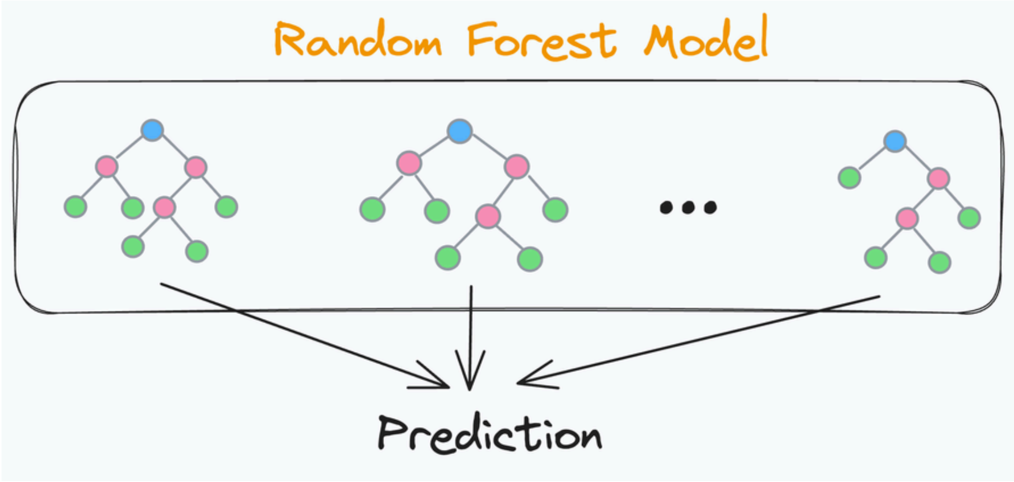
Modeling Approach

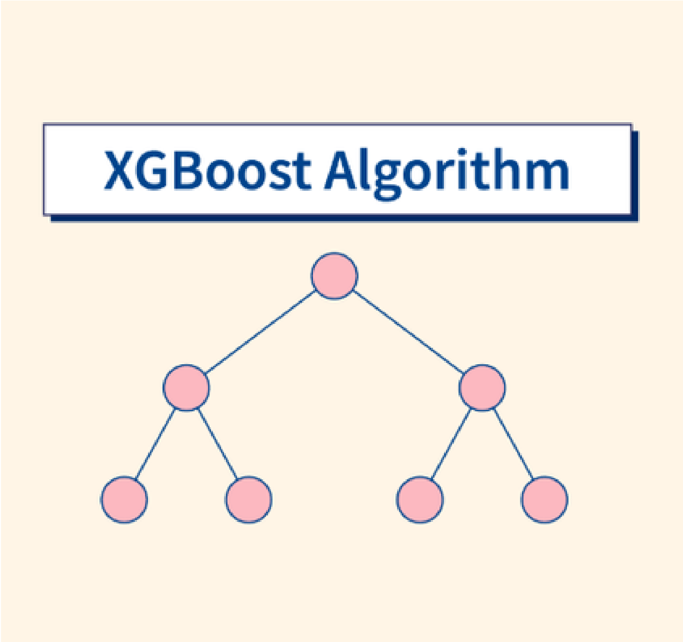
Models implemented:

- Random Forest Classifier (main model)
- XGBoost

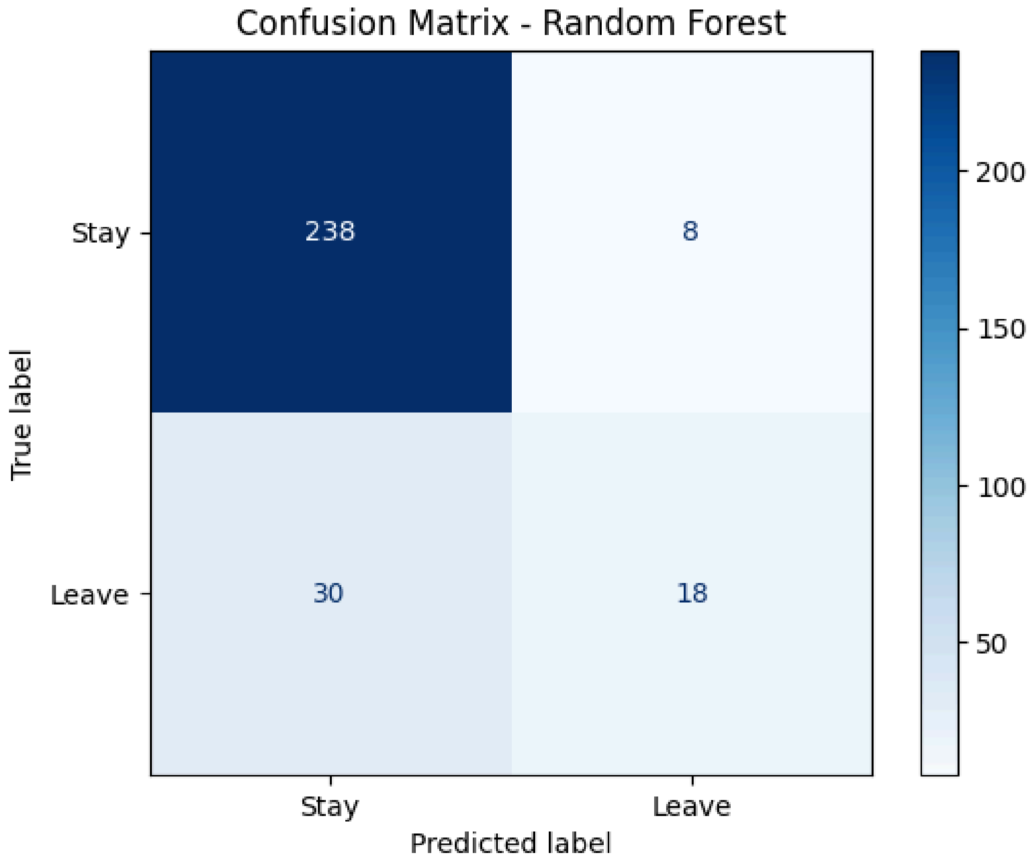
Evaluation Metrics:

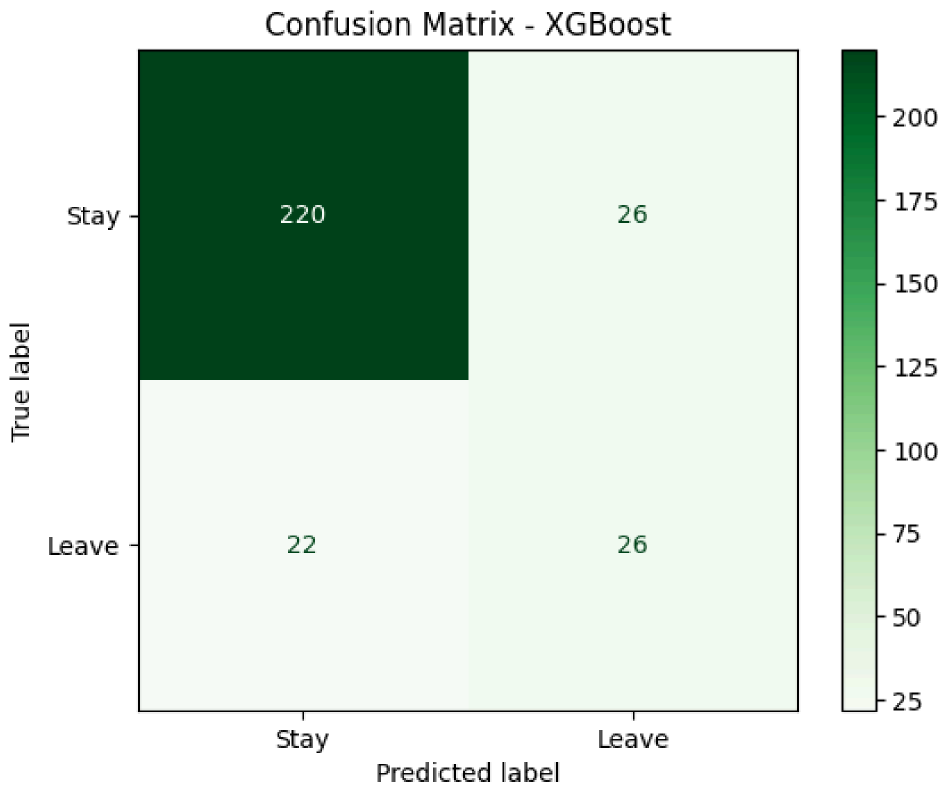
- Accuracy
- Precision
- Recall
- F1 Score
- Confusion Matrix
- Feature Importance





Model Performance





RESULTS OF RANDOM FOREST:
ROC-AUC Score: 0.823

Detailed Report:
Accuracy: 0.8707482993197279
Precision: 0.6923076923076923
Recall: 0.375
F1 Score: 0.4864864864864865

RESULTS OF XGBOOST
ROC-AUC Score: 0.820

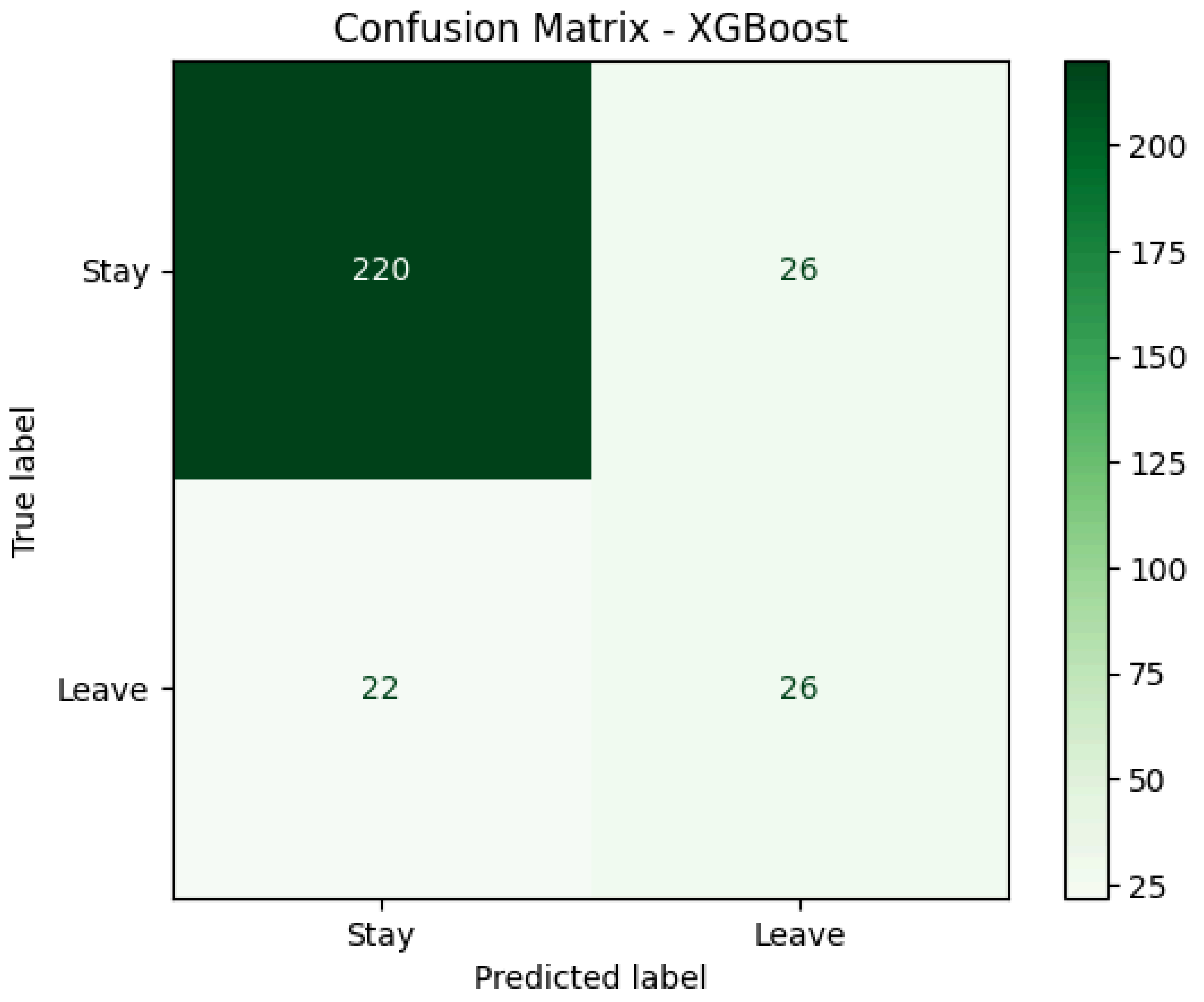
Detailed Report:
Accuracy: 0.8367346938775511
Precision: 0.5
Recall: 0.5416666666666666
F1 Score: 0.52

Summary

- Logistic Regression provides baseline interpretability
- Random Forest performs better overall and captures non-linear patterns
- Improved recall on “Attrition = Yes” category

Highlight

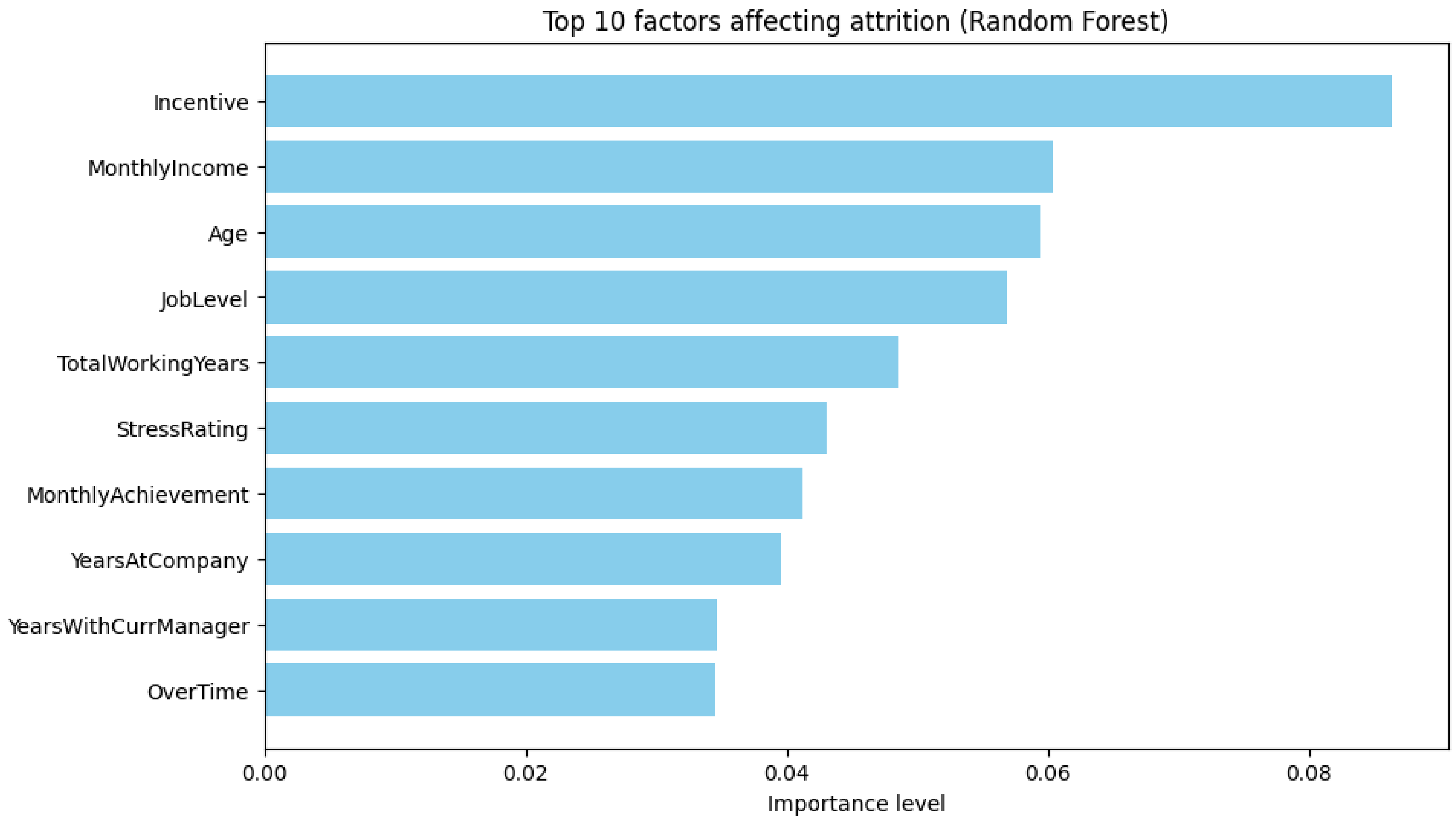
False Negatives = employees who actually left but were predicted to stay
→ The company should minimize this group, as they represent unexpected resignations.

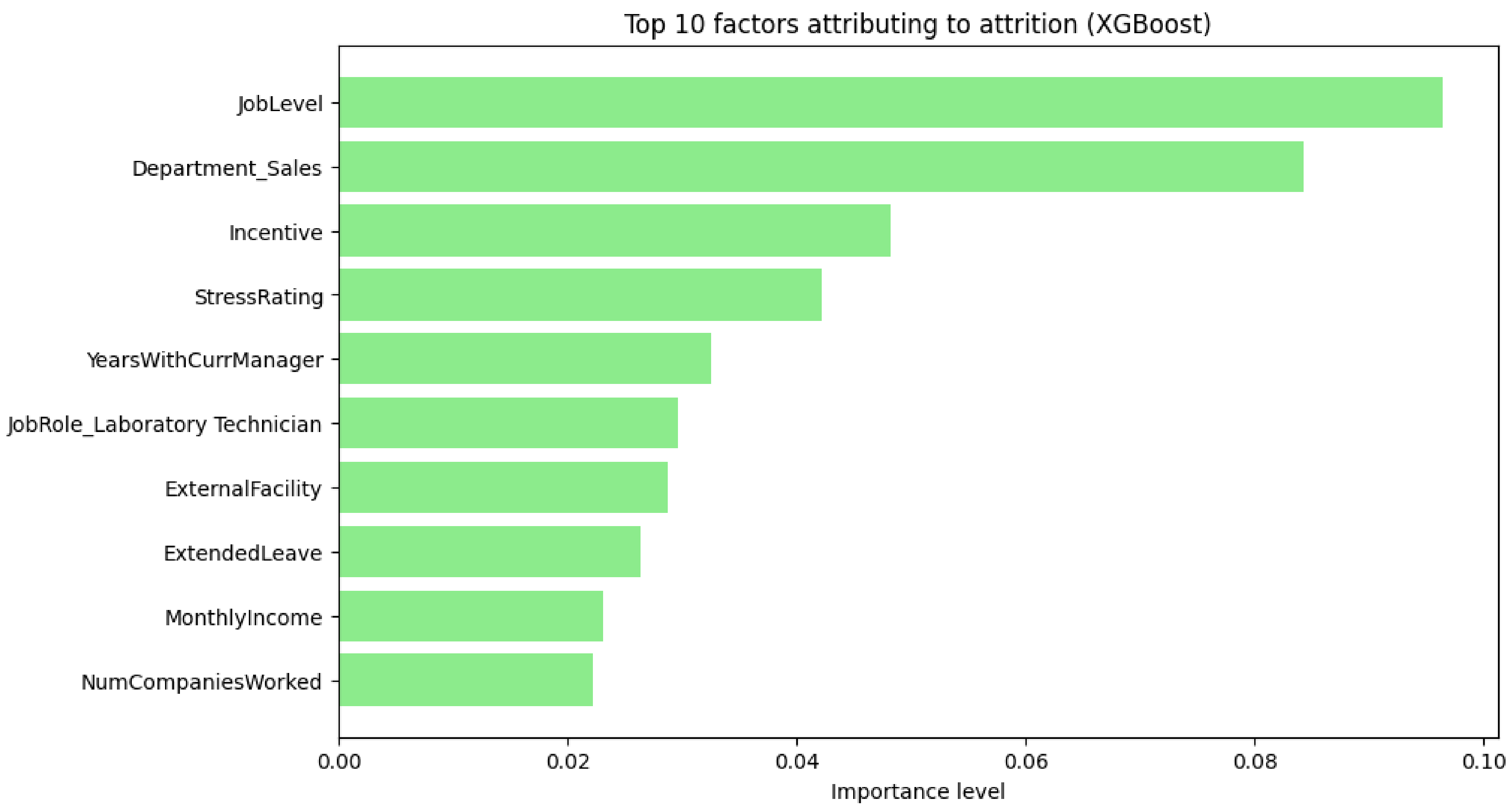


Feature Importance

The top contributing factors include:

1. Incentive: The strongest predictor in both models
2. Monthly Income & Job Level: Compensation and career level play critical roles.
3. YearsWithCurrManager: Relationship with the direct manager strongly affects retention.
4. Stress & OverTime: Workload and job pressure significantly impact attrition risk.





Key Findings

Early-Career Employees Are the Most Vulnerable

Young, single, low-tenure employees → higher attrition risk

Sales roles, frequent travel, and low satisfaction levels are strong signals

Income and career progression play major roles

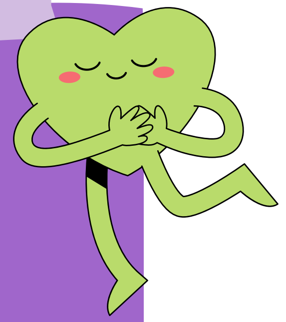
Interns have notably low attrition → strong potential talent pipeline

Stress is not the main driver; satisfaction matters more

Strategic Proposals to Reduce Attrition

Strengthen Onboarding & Early-Career Support (High Priority)

Optimize High-Pressure Roles (Sales & Frequent Travel)



Improve Work Environment & Managerial Relationships

Actions:

- *Launch “First 100 Days Success Program” for new hires*
- *Assign mentors for first-year employees*
- *Provide clear 12–24 month career progression roadmaps*
- *Conduct quarterly check-ins for early-tenure staff*

- *Reduce mandatory travel or rotate travel assignments*
- *Provide travel stipends & hybrid meeting solutions*
- *Implement Sales wellness support (coaching, training)*
- *Evaluate workload distribution in Sales Representative roles*

- *Improve team culture through leadership training*
- *Introduce anonymous monthly satisfaction surveys*
- *Implement quick-win fixes (equipment, workspace, process clarity)*
- *Reward team leaders with low turnover*