IBM HR Analytics
Employee Attrition & Performance

Detailed Project Report

Department	Human Resources
Contributor/s	Swapnil Tayde

Tools used	• MS Excel
	• Python
	  ○ Pandas
	  ○ NumPy
	  ○ Matplotlib
	  ○ Seaborn
	  ○ SciPy
	  ○ scikit-learn
	• Jupyter Notebook
	• Power BI

Project focus	• Attrition diagnosis • Driver ranking • Retention strategy

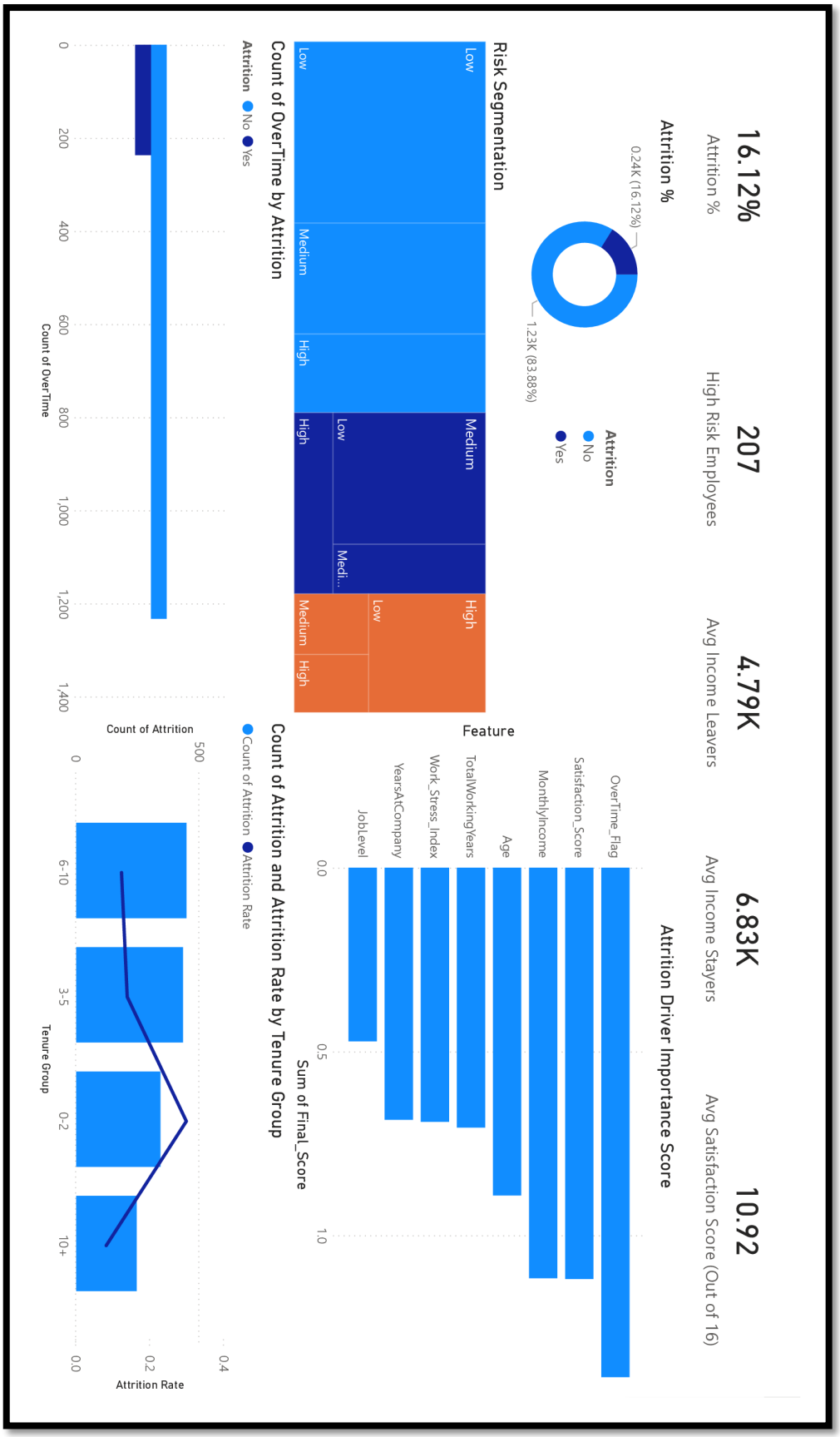

Figure 0. Power BI dashboard preview used to communicate the attrition story.

Executive Summary

Project snapshot

- The dataset contains 1,470 employee records with 35 original fields and no missing values or duplicates.
- Overall attrition is 237 employees, equal to 16.1% of the workforce.
- The strongest risk signals are overtime, low satisfaction, lower income, early tenure, and younger age bands.

This analysis examines IBM HR attrition patterns through a combination of exploratory analysis, statistical testing, feature engineering, and machine learning. The purpose is not only to describe who is leaving, but to isolate the variables that consistently differentiate employees who stay from those who exit.

The evidence shows a clear concentration of attrition among employees who work overtime, earn less, have shorter tenure, and report weaker satisfaction outcomes. Attrition is also materially higher in Sales-related roles, among single employees, and among those who travel frequently.

From a modeling standpoint, the random forest classifier reached 0.85 accuracy on the test set, but recall for the attrition class remained low because the target is imbalanced. That makes the analysis especially useful as a decision-support layer: it highlights the highest-risk segments even where raw prediction accuracy is not the whole story.

Metric	Value	Interpretation
Total employees	1,470	Analysis population
Attrition count	237	Employees who left
Attrition rate	16.1%	Baseline churn level
Overtime rate	28.3%	Operational pressure indicator
High-risk combined segment	33 employees	Low satisfaction + high stress + low tenure + low income

Key message: attrition is not random. It clusters in identifiable workforce segments that can be targeted through compensation review, workload balancing, manager intervention, and career-path stabilization.

Problem Statement & Objectives

The organization needs a repeatable, evidence-based way to understand employee attrition and performance risk. A descriptive report alone is not sufficient; leadership needs to know which variables matter most, how strongly they matter, and what actions are likely to reduce churn.

The project was built around four practical objectives:

- Measure the overall attrition profile and identify where turnover is concentrated.
- Test whether important workforce attributes differ significantly between employees who stay and employees who leave.
- Engineer interpretable features that capture stress, satisfaction, commute burden, and career progression.
- Rank the strongest attrition drivers and translate them into business actions for retention planning.

The analysis is intended for HR stakeholders, business leaders, and analysts who need a compact view of risk drivers without losing the statistical grounding behind the conclusions.

Scope of the analysis

- Population: 1,470 employees from the IBM HR attrition dataset.
- Target variable: Attrition (Yes/No).
- Methods: univariate analysis, bivariate analysis, chi-square tests, t-tests, Cohen's d, logistic regression, random forest, and combined driver ranking.

Data Cleaning & Preparation

The dataset was already structurally clean: there were no missing values and no duplicate rows. The preparation step therefore focused on schema simplification and analysis-ready transformation rather than heavy imputation or record repair.

What was done

- Removed constant columns that add no analytical value: EmployeeCount, Over18, and StandardHours.
- Converted categorical variables to category dtype to improve memory usage and keep the schema explicit.
- Converted Attrition into a binary target field for modeling and risk analysis.
- Validated the dataset for structural integrity before feature engineering and testing.

Data quality checks

Check	Result	Implication
Missing values	0	No imputation required
Duplicate rows	0	No deduplication required
Constant columns removed	3	Schema cleanup
Analytical target	Attrition	Binary classification

The result is a stable base dataset that supports both exploratory work and supervised learning. Because the source data was clean, the strongest analytical lift came from feature engineering and segmentation rather than from repair tasks.

Full EDA

1. Attrition baseline and workforce profile

Out of 1,470 employees, 1233 stayed and 237 left. The baseline attrition rate is 16.1%, which is substantial enough to warrant targeted retention actions.

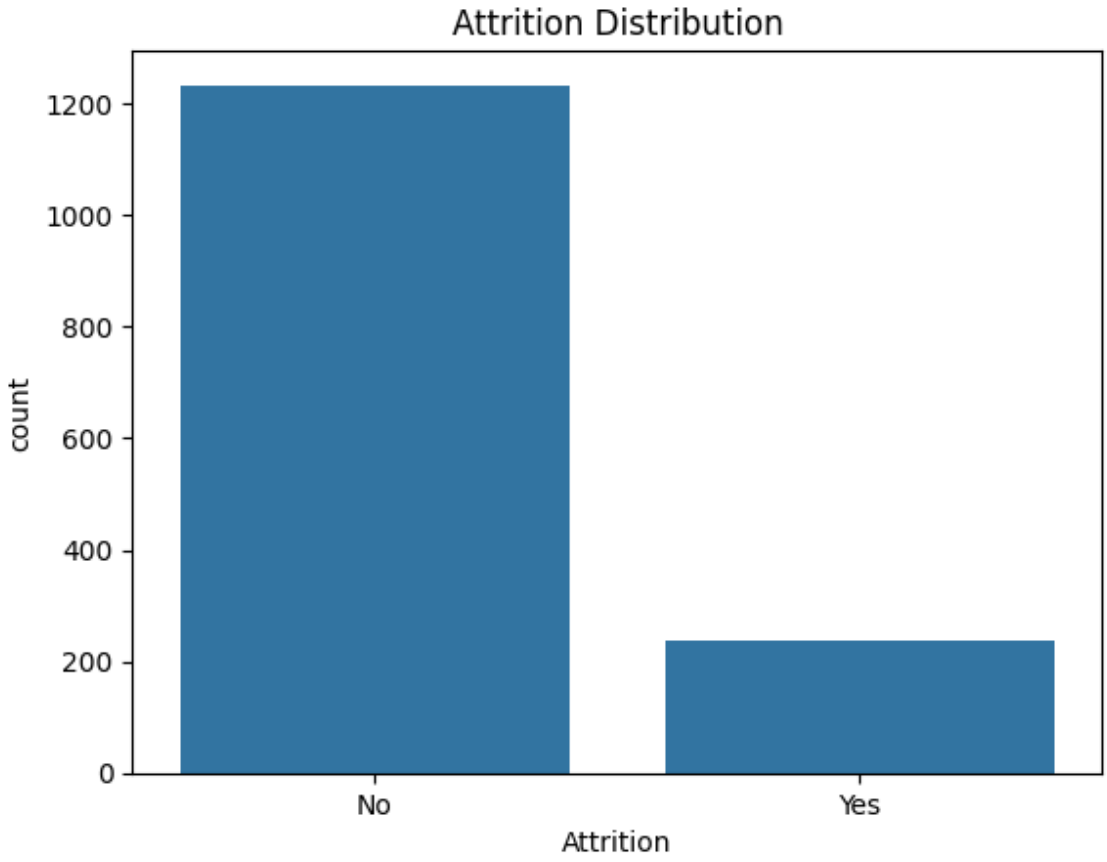

Figure 1. Attrition distribution across the full workforce.

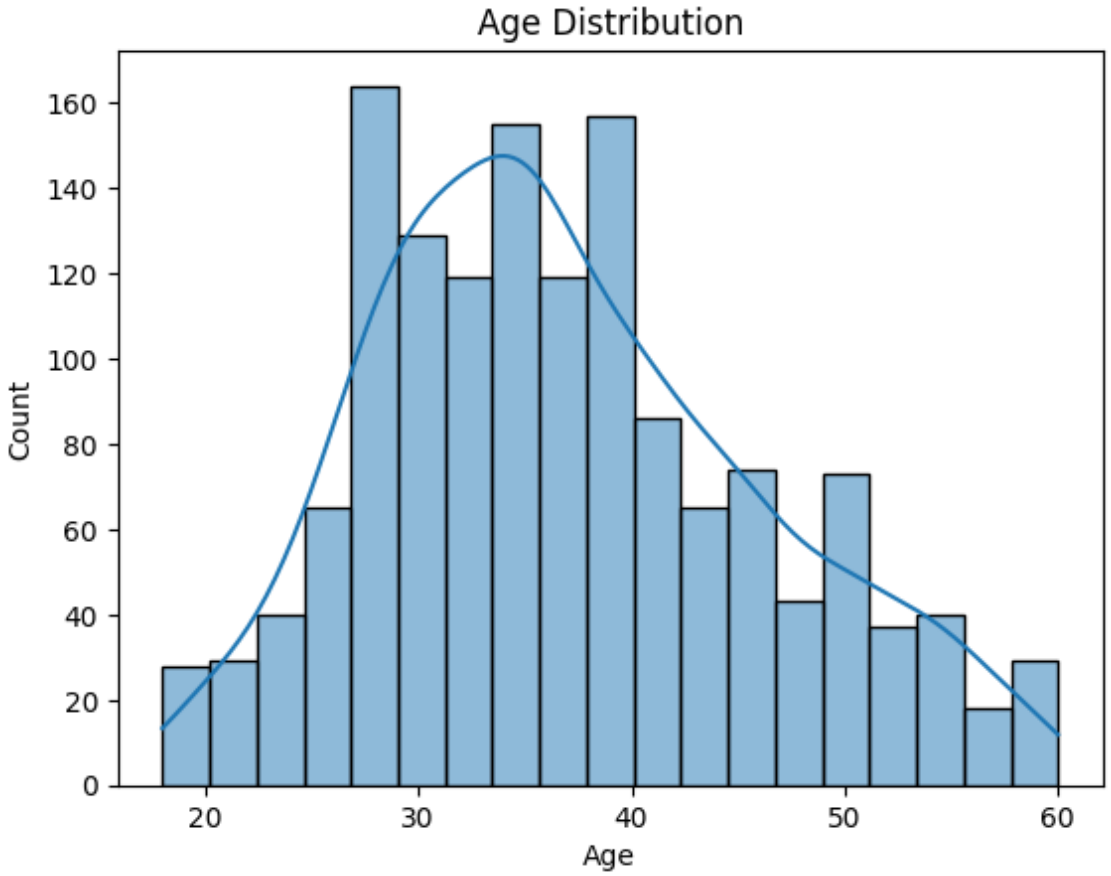

Figure 2. Age distribution, showing a workforce centered in the 26-35 and 36-45 bands.

2. Department and education mix

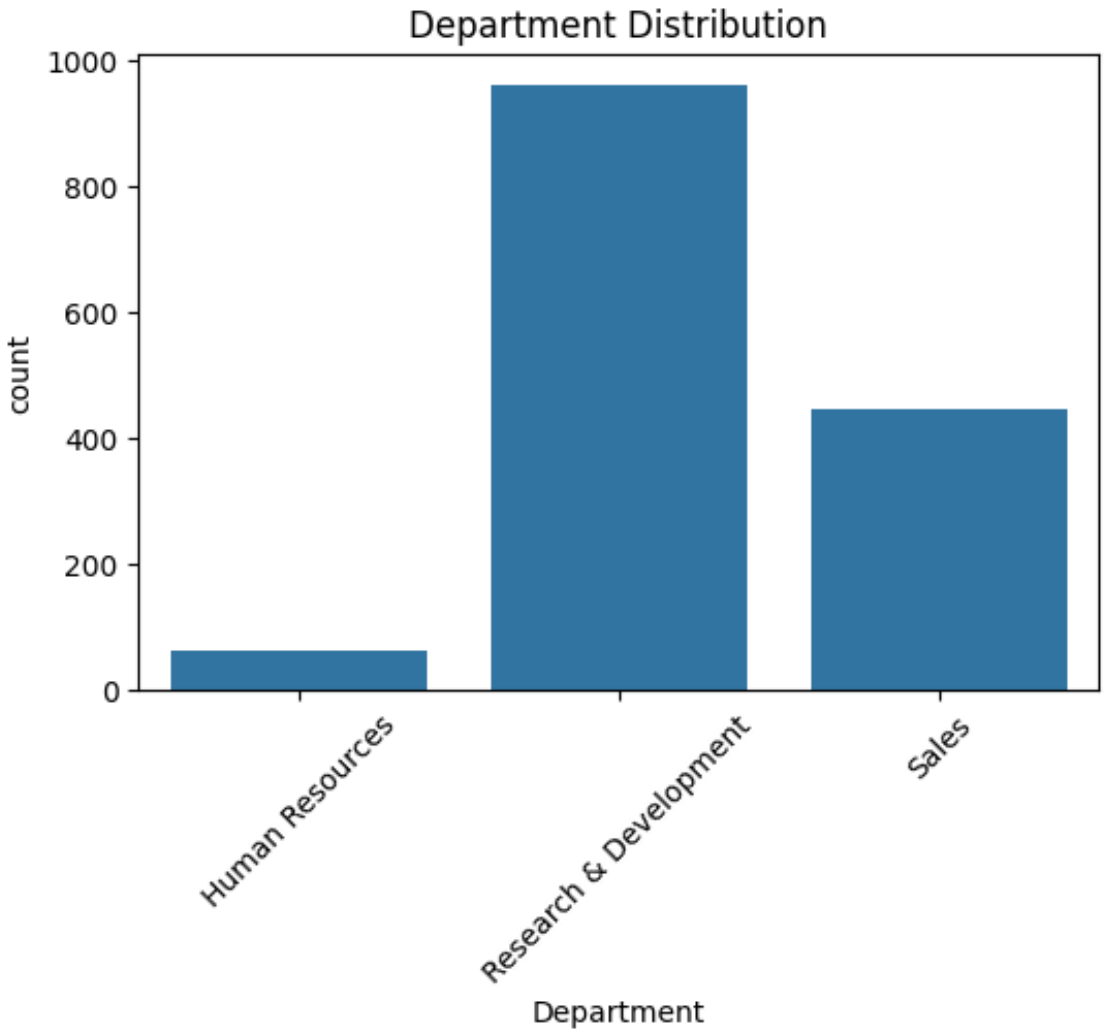

Figure 3. Department distribution, with Research & Development holding the largest share.

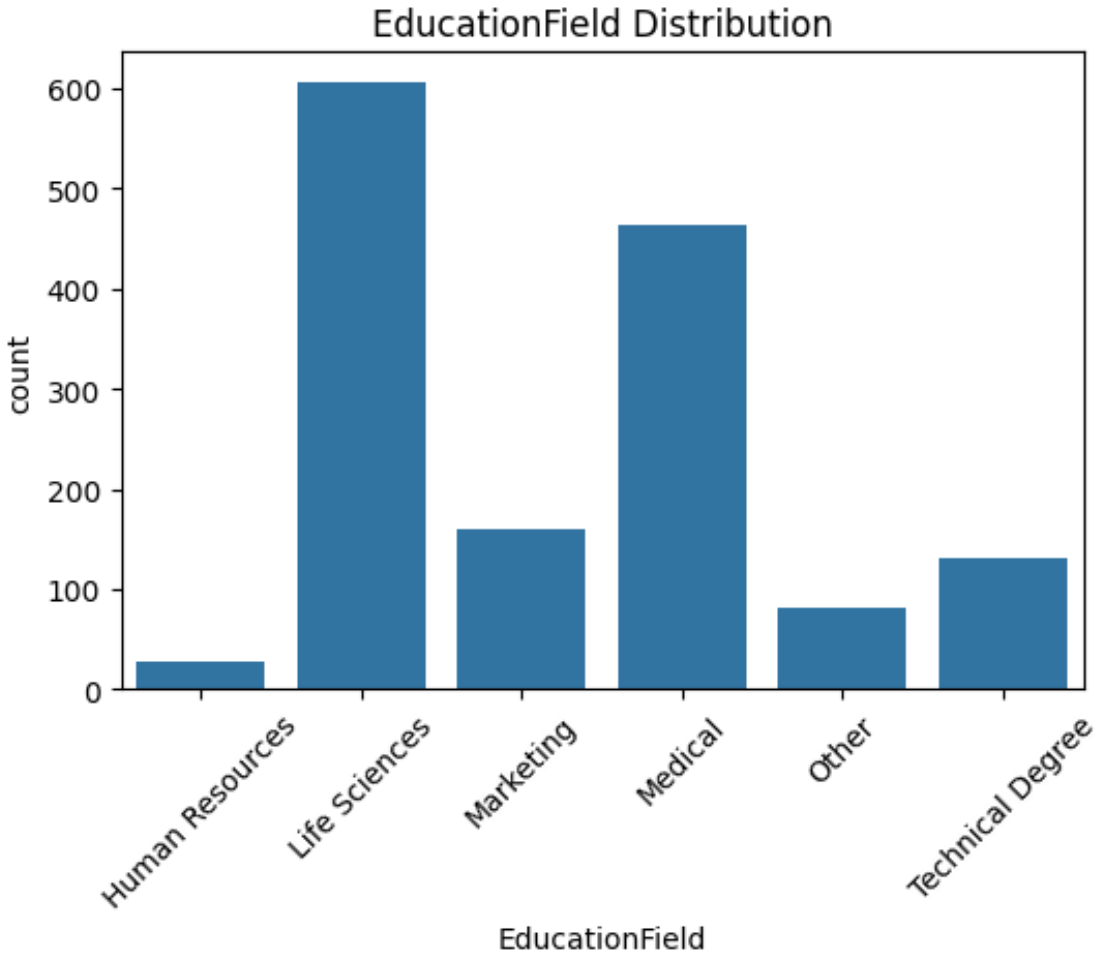

Figure 4. Education field distribution, used to understand talent background mix.

Full EDA

3. Role, travel, and overtime patterns

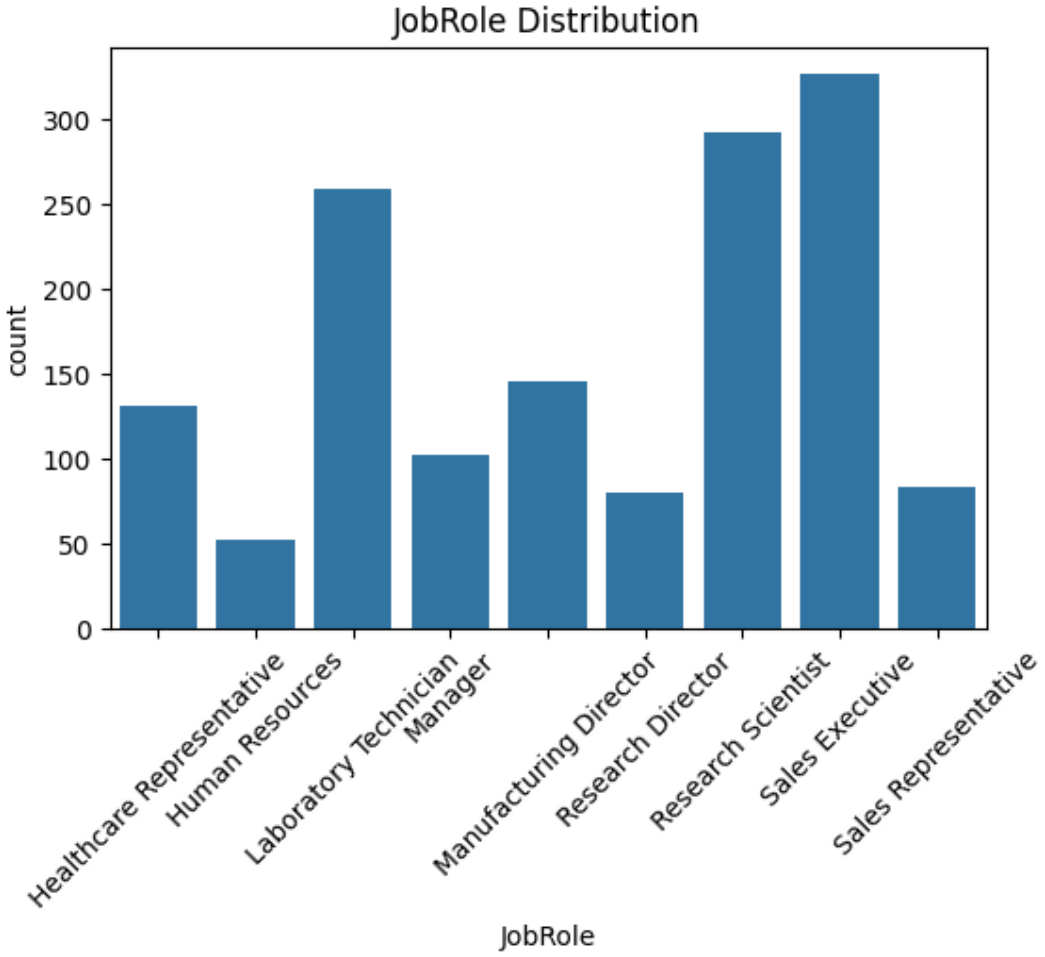

Figure 5. Job role distribution.

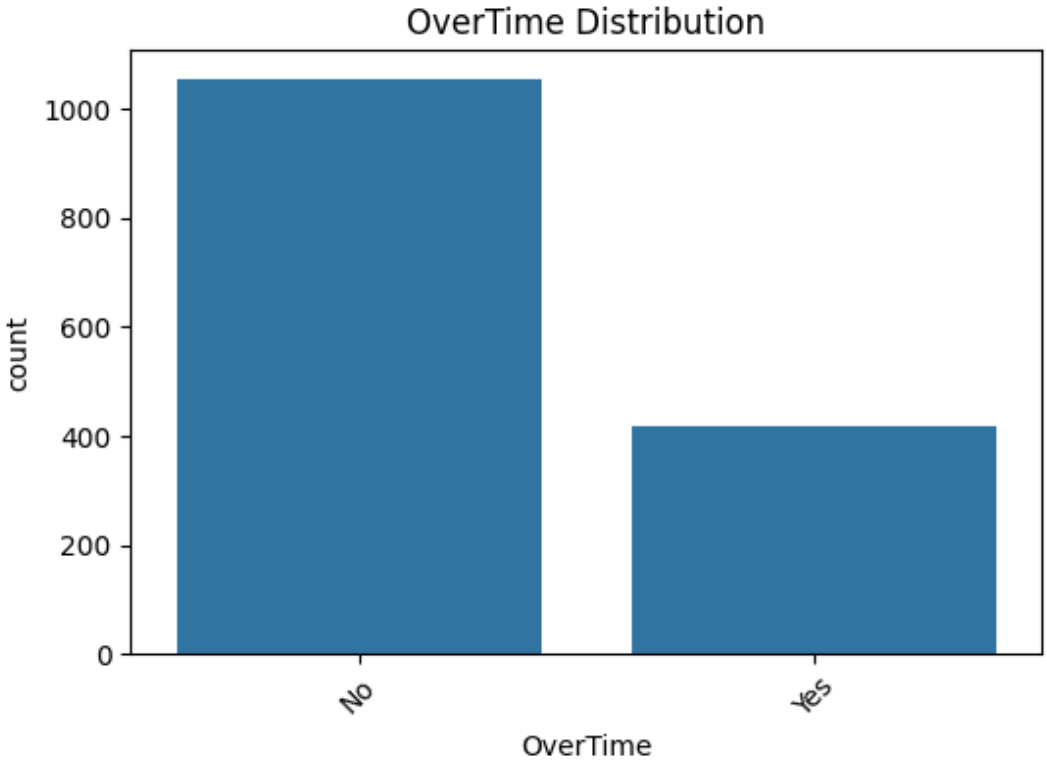

Figure 6. Overtime distribution, a critical attrition signal.

Figure 7. Business travel distribution.

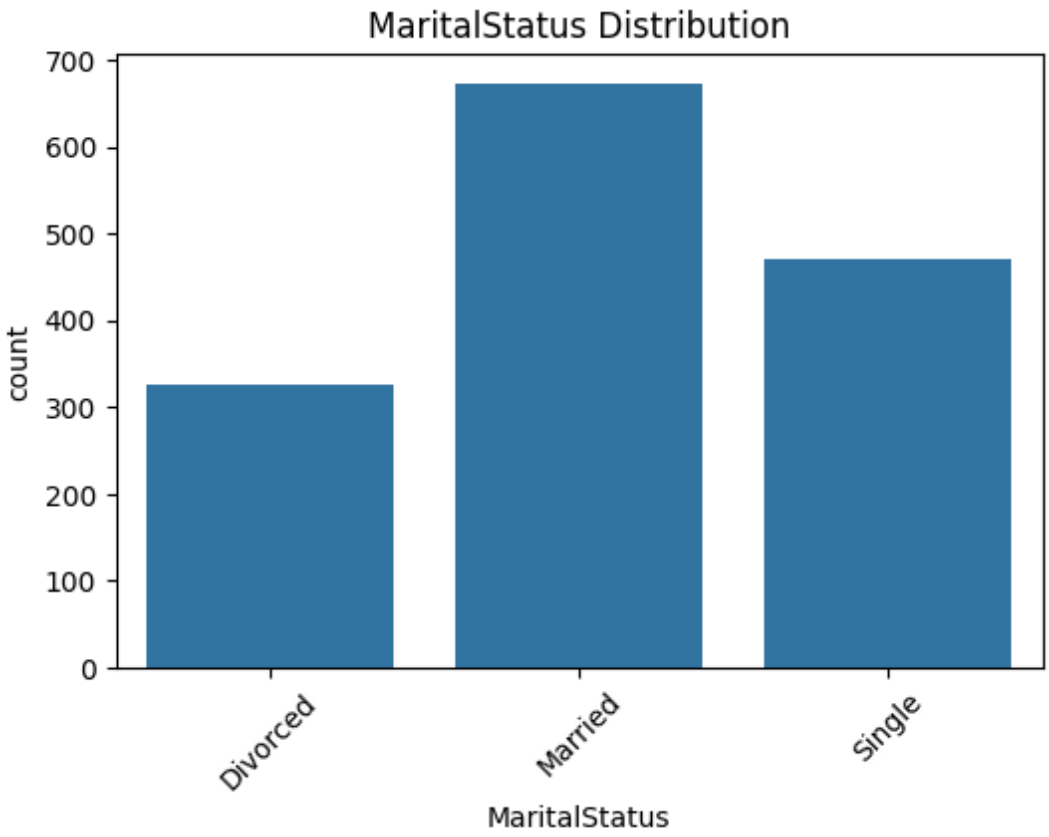

Figure 8. Marital status distribution.

The EDA already points to an operational pattern: attrition is more common in roles and work arrangements that carry higher travel intensity, overtime, or customer-facing pressure.

Full EDA

4. Compensation, tenure, and correlation structure

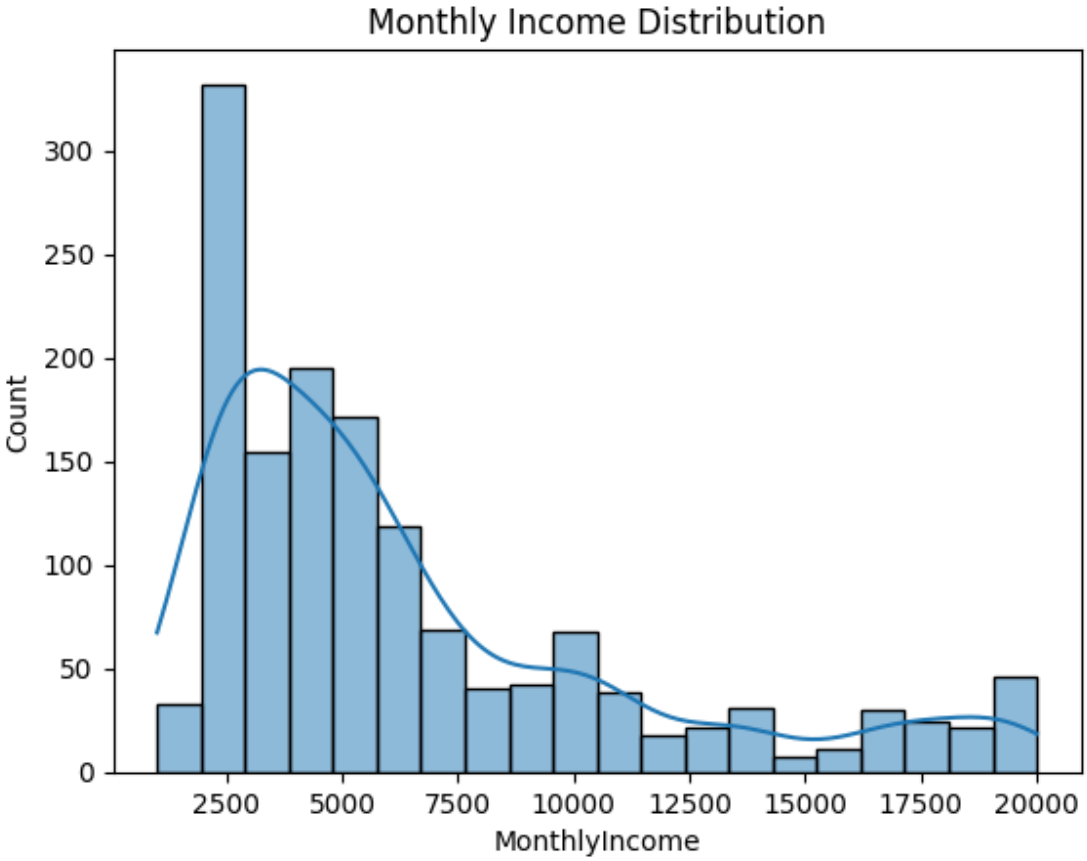

Figure 9. Monthly income distribution.

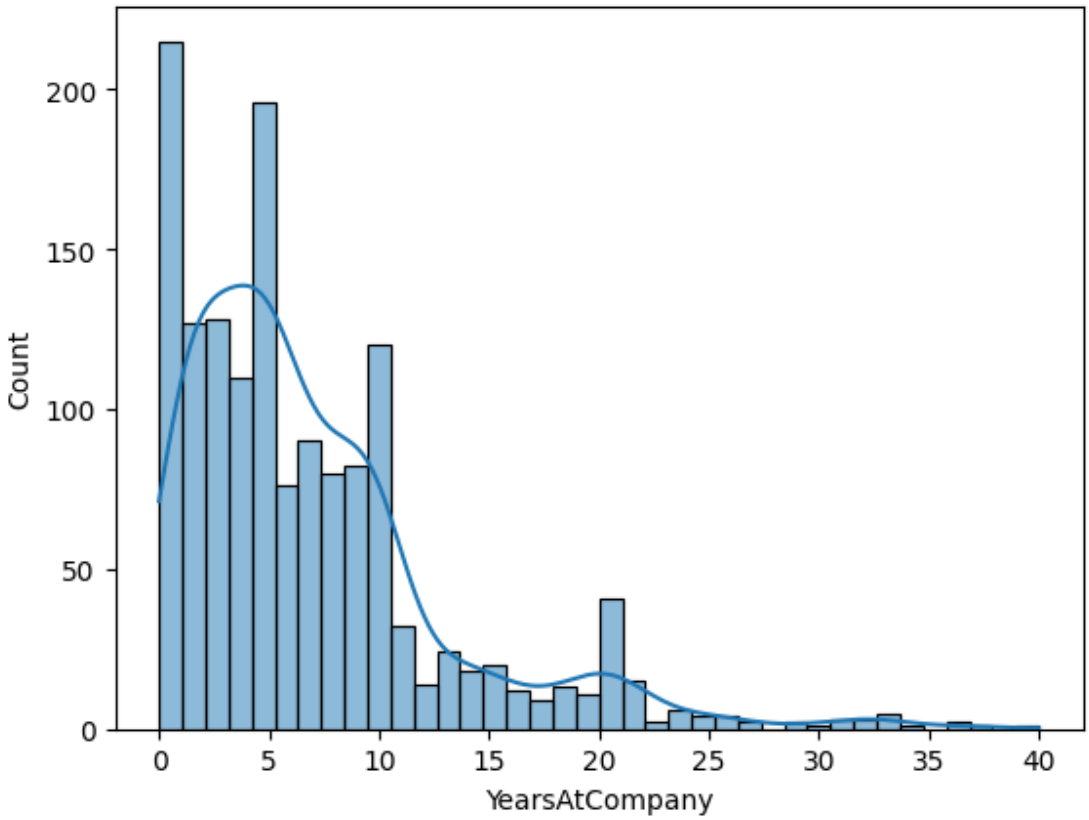

Figure 10. Tenure distribution, showing many employees in the early to mid-tenure range.

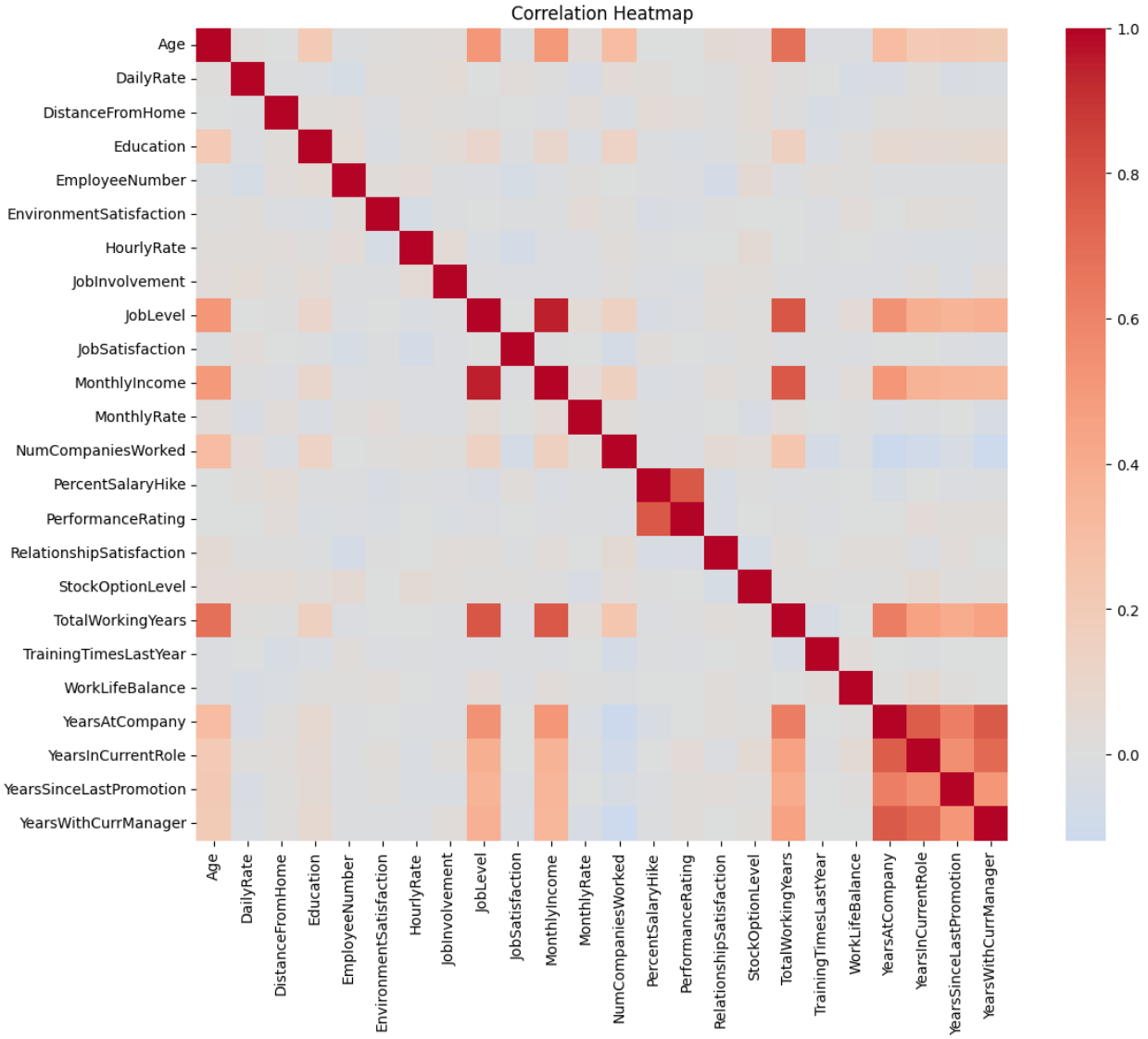

Figure 11. Correlation heatmap for the numeric feature set.

EDA interpretation

- Income and tenure are among the most informative numeric dimensions for attrition review.
- The correlation matrix does not indicate extreme multicollinearity across the core engineered variables.
- Outlier checks on MonthlyIncome and TotalWorkingYears are useful for spotting a small number of unusually positioned employees.

Feature Engineering

Feature engineering was used to convert raw HR fields into business-relevant signals. This step improved interpretability and made the later statistical and machine-learning stages more actionable.

Engineered feature	Formula / logic	Why it matters
OverTime_Flag	OverTime mapped to 1/0	Binary workload indicator
Work_Stress_Index	JobInvolvement + OverTime_Flag + (5 - WorkLifeBalance)	Higher score = higher stress
Satisfaction_Score	JobSatisfaction + EnvironmentSatisfaction + RelationshipSatisfaction + WorkLifeBalance	Composite satisfaction view
Commute_Intensity	DistanceFromHome x OverTime_Flag	Commute burden under overtime
Career_Growth_Rate	JobLevel / (TotalWorkingYears + 1)	Career progression proxy
Age_Group	Binned age bands	Lifecycle segmentation
Experience_Group	Binned tenure bands	Experience segmentation

The two most important composites are Work_Stress_Index and Satisfaction_Score. They summarize multiple raw variables into interpretable scores that are easier to compare across workforce segments and easier to explain to leadership.

Statistical Analysis

The statistical stage tested whether attrition-related differences were likely to be real rather than visual artifacts. Categorical variables were evaluated with chi-square tests, while numeric variables were compared using independent-samples t-tests and Cohen's d.

Variable	p-value	Effect / interpretation
OverTime	< 0.00001	Significant association
Department	0.00453	Significant association
JobRole	< 0.00001	Significant association
MaritalStatus	< 0.00001	Significant association
BusinessTravel	< 0.00001	Significant association
MonthlyIncome	< 0.00001	d = -0.479
YearsAtCompany	< 0.00001	d = -0.372
Work_Stress_Index	0.02853	d = 0.150
Satisfaction_Score	< 0.00001	d = -0.427

Selected mean comparisons are directionally consistent with the plots:

- Employees who left had lower mean monthly income than those who stayed.
- Employees who left had shorter average tenure and lower total working years.
- Employees who left reported lower composite satisfaction scores.
- Work stress was slightly higher among leavers, though the effect size was smaller than the satisfaction and income gaps.

Overall, the statistical tests support the visual story: attrition is strongly linked to compensation, tenure, satisfaction, and workload patterns.

Machine Learning Model Explanation

Two models were used for driver analysis. Logistic regression supplied directionality and coefficient-based interpretability. Random forest captured nonlinear effects and relative feature importance. The goal was explanation, not just prediction.

Model / step	Purpose	Value in this project
Logistic Regression	Interpretable coefficients	Shows whether a feature pushes attrition risk up or down
Random Forest	Nonlinear ensemble	Captures interaction effects and ranks predictive power
Train/test split	80/20 with stratification	Preserves class balance in evaluation
Target encoding	Attrition_Flag (0/1)	Standard binary classification target

Model evaluation

Metric	Score	Meaning
Accuracy	0.85	Overall correct classifications
Class 0 precision	0.88	Most non-attrition cases were correct
Class 0 recall	0.96	Very strong retention-class detection
Class 1 precision	0.31	Limited precision for leavers
Class 1 recall	0.10	Attrition cases remain difficult to capture
Class 1 F1-score	0.15	Low minority-class balance

The performance profile is typical of imbalanced workforce data. A high headline accuracy can hide weak detection of the minority class, so the classification report must be read alongside the attrition rate itself.

Driver Ranking

The final driver ranking combines the standardized logistic regression coefficient strength with random forest importance. This hybrid approach balances directional interpretability with predictive relevance.

Rank	Driver	Interpretation
1	OverTime_Flag	Strongest workload signal
2	Satisfaction_Score	Composite satisfaction pressure
3	MonthlyIncome	Compensation disadvantage
4	Age	Lifecycle-related risk
5	TotalWorkingYears	Experience depth
6	Work_Stress_Index	Work pressure proxy
7	YearsAtCompany	Tenure stability
8	JobLevel	Career position

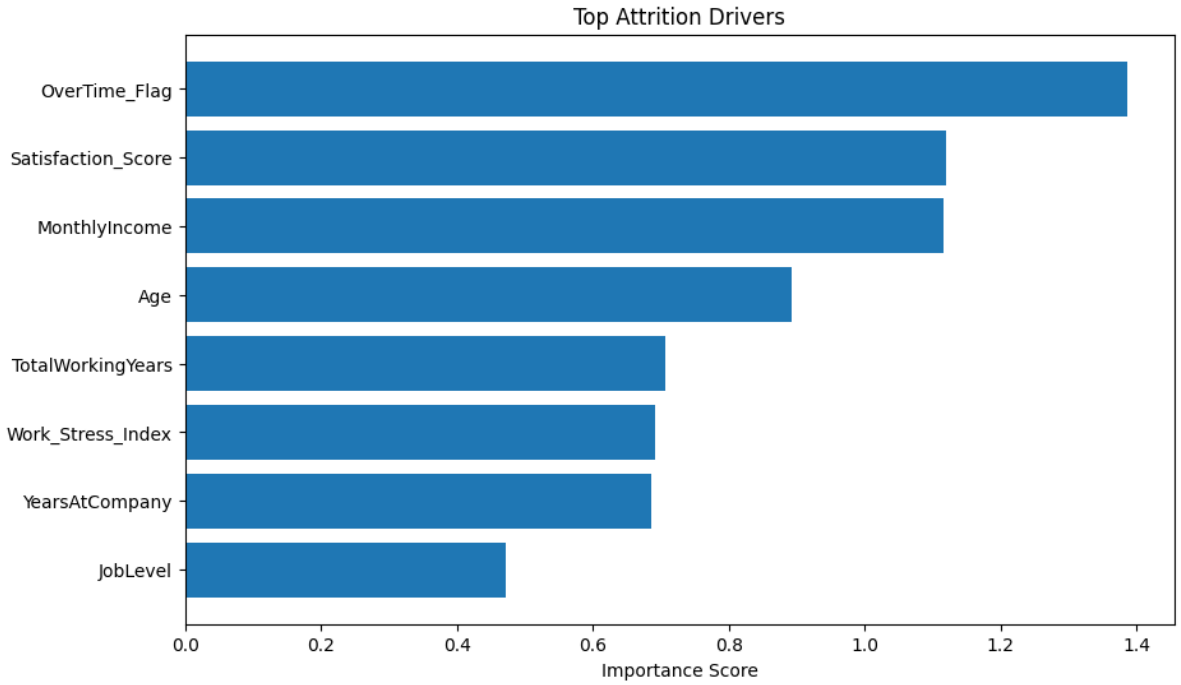

Figure 12. Top attrition drivers ranked by combined model score.

The ranking suggests a practical hierarchy: first address overtime and satisfaction, then review compensation and tenure-related pathways, and finally refine management interventions for early-career employees.

Business Insights

- Employees working overtime show 30.5% attrition versus 10.4% for employees without overtime.
- Sales Representative attrition is the highest among job roles at 39.8%.
- Single employees have a materially higher attrition rate (25.5%) than married or divorced employees.
- Employees in the 18-25 age band and the 0-5 years experience band show the strongest early-career churn.

  The business interpretation is straightforward: the attrition problem is concentrated in a few manageable segments rather than spread uniformly across the workforce. That means retention actions can be narrower, more cost-effective, and easier to monitor.

High-risk segment view

Segment	Employees	Attrition rate
Low satisfaction segment	601	21.6%
High stress segment	583	18.5%
Low tenure segment	470	26.0%
Low income segment	369	29.3%
Combined high-risk group	33	45.5%

Strategic Recommendations

- Reduce avoidable overtime through workload balancing, shift redesign, and manager-level overtime approvals.
- Review compensation bands for lower-income and early-tenure employees, especially in Sales and Sales Representative roles.
- Strengthen onboarding and first-two-year retention programs to protect the 0-5 year experience group.
- Use satisfaction and stress scores as early-warning indicators in HR dashboards for monthly review.
- Introduce manager interventions for single, frequently traveling, and early-career employees in higher-risk departments.

Executive takeaways

This project shows that attrition can be explained with a relatively small number of high-value signals. Overtime, satisfaction, income, age, and tenure carry the most decision value. The most effective retention strategy is therefore not a blanket policy, but a targeted response by segment and risk level.

The analysis can be extended further by adding cost-of-turnover estimation, department-level manager effects, and time-based trend analysis. For now, the report provides a solid decision layer for HR planning and retention prioritization.