Employee Attrition Analysis Report

Data Overview and Preparation

Dataset Source and Structure

The analysis uses the public IBM HR Analytics Employee Attrition dataset from Kaggle, comprising 1,470 employee records with 35 features, including demographics, job details, compensation, and tenure. The target variable, Attrition, is binary ("Yes"/"No"), with a severe imbalance: 1,233 retained (84%) vs. 237 attrited (16%).

Data Cleaning and Preprocessing

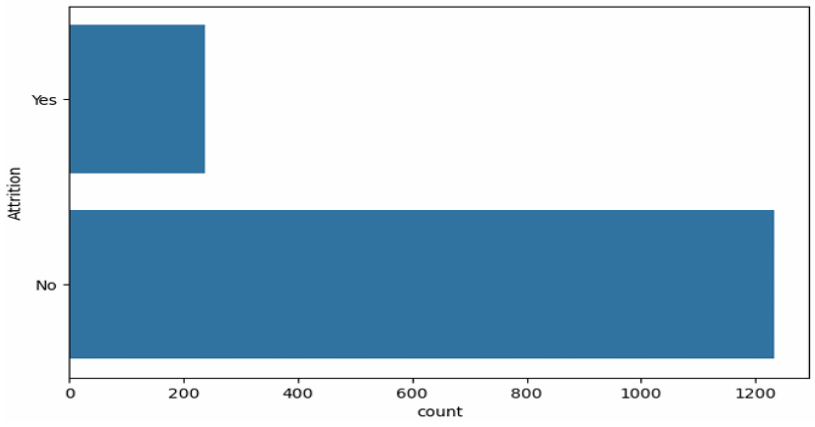
- Null and Duplicate Checks: No missing values or duplicates were found.
- Descriptive Statistics: Average age is 37, monthly income $6,503, and years at company 7. Attrition correlates with higher distance from home and overtime.
- Feature Selection: Dropped constant columns to reduce to 32 features. No low-variance features post-split.
- Encoding: Label-encoded binary features (Attrition: No=0/Yes=1, Gender: Male=0/Female=1, OverTime: No=0/Yes=1). One-hot encoded categorical features (e.g., Department, JobRole), expanding to 46 columns.
- Handling Imbalance: Applied RandomOverSampler to balance classes, resulting in 2,466 samples.
- Scaling and Split: Standardized numeric features. Split into 80% train and 20% test.

VarianceThreshold confirmed no constant features in training data, ensuring robust input for modeling.

Exploratory Data Analysis

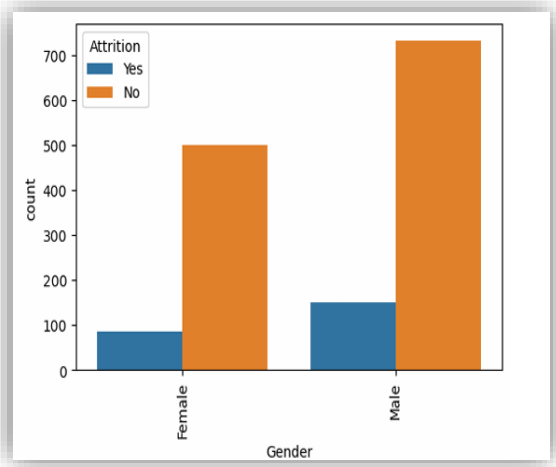
Attrition Distribution

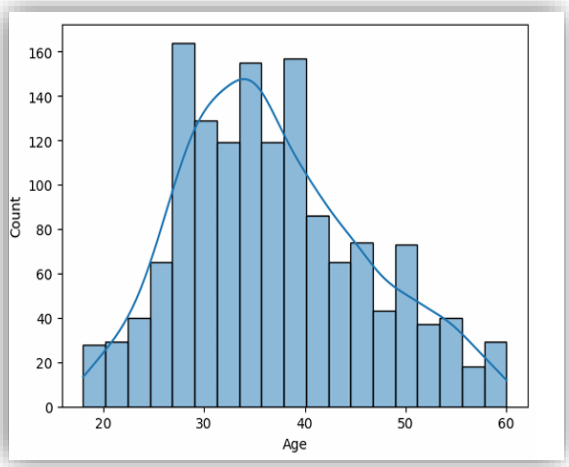
- Overall Rate: 16% (237 cases), below the 2025 U.S. average voluntary turnover of 20% but moderate for tech/R&D sectors (20%).
- Imbalance: 84% retained, necessitating oversampling for modeling.

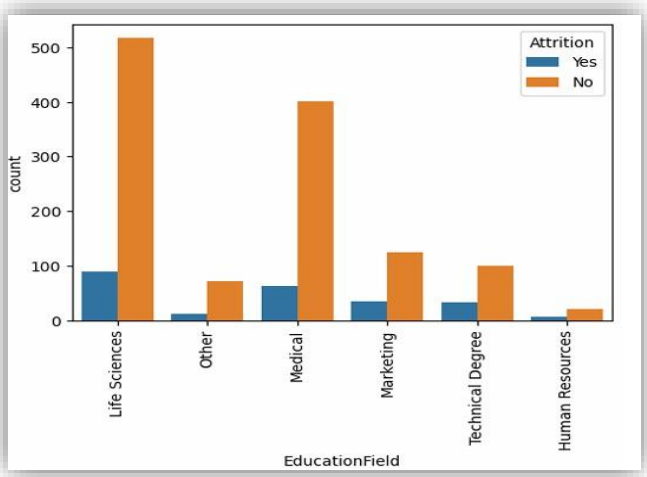


Demographic Trends

- Gender: Males (60% of workforce) have higher attrition (63% of cases), indicating potential gender-specific issues.
- Age: Peaks in 30-50 group (mean 37), with under-30s showing early turnover—common in entry roles per 2025 benchmarks.
- Education Field: STEM-dominant (Life Sciences 38%, Medical 27%), with higher attrition in these areas due to external demand.

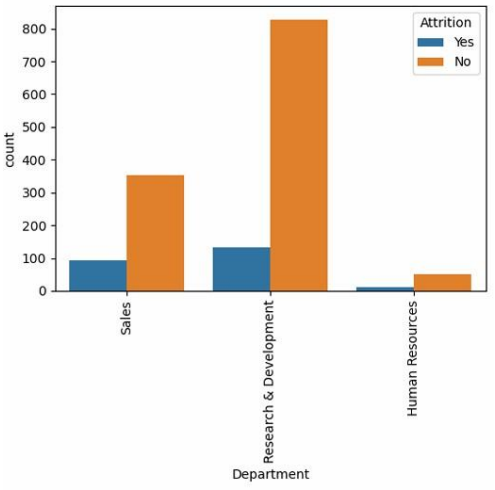


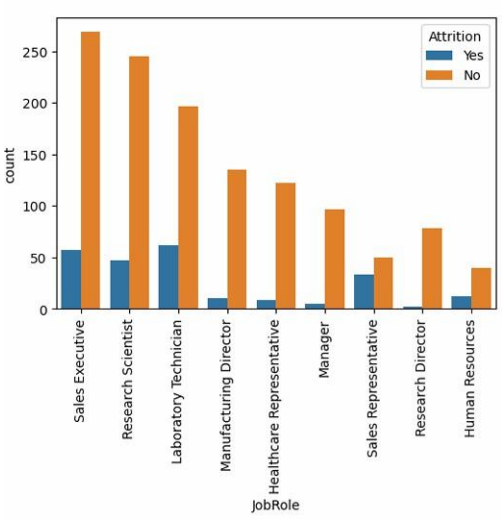


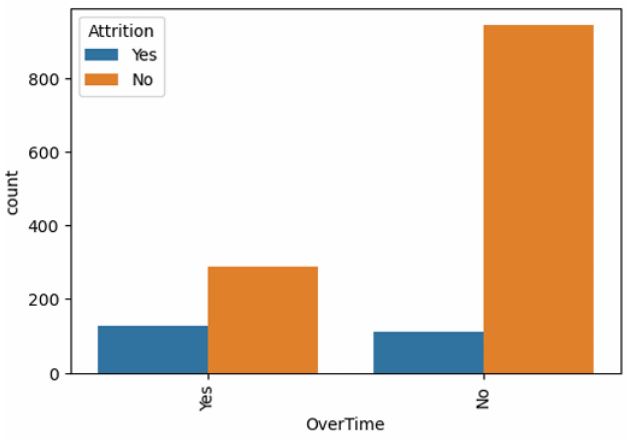


Job and Tenure Trends

- Job Roles: Highest in Laboratory Technician (62), Sales Executive (57), Research Scientist (47). Frontline R&D/Sales roles align with elevated turnover in competitive industries.
- Departments: Research & Development (65% of workforce) drives most attrition, followed by Sales.
- Tenure: Attrition peaks in first year, declining after 10 years, proxying onboarding inefficiencies.
- Overtime and Satisfaction: 28% work overtime, correlating with attrition. JobSatisfaction averages ~2.7 (on 1-4 scale), below U.S. averages of 50-62% satisfaction.
- Histograms showed normal distributions for age and income, with countplots highlighting attrition spikes in single employees and frequent travelers.







Model Building and Performance

Methodology

A Random Forest Classifier model was selected for its interpretability in forecasting attrition risk based on factors like tenure, performance, compensation, and work-life elements.

After preprocessing:

- Trained on oversampled, scaled data.
- No hyperparameters tuned (default settings).

Results

- Predictions: Model outputs binary classifications (0=retained, 1=attrited).
- Accuracy: 96% on test set.
- Classification Report:

Class	Precision	Recall	F1-Score	Support
0 (Retained)	0.94	0.97	0.96	244
1 (Attrited)	0.97	0.94	0.96	250
Macro Avg	0.96	0.96	0.96	494
Weighted Avg	0.96	0.96	0.96	494

- Confusion Matrix: Balanced errors (e.g., 8 false negatives, 14 false positives), suitable for risk forecasting.
- Model Saving: Exported via joblib for deployment.

This performance indicates the model effectively captures attrition drivers, outperforming random guessing (50%) and supporting proactive interventions.

Insights and Trends

- High-Risk Groups: Early-tenure employees (<5 years), overtime workers, and R&D/Sales roles face elevated risks, mirroring 2025 trends in tech (20%+ attrition). Factors like low job involvement and high distance from home amplify this.
- Compensation and Performance: Lower salary hikes correlate with attrition, while high performers stay longer.
- Diversity Implications: Male and STEM-heavy attrition suggests retention gaps, with opportunities to diversify non-technical roles.
- Benchmark Comparison: The 16% rate is below retail (26.7%) but near healthcare (18-20%), indicating moderate stability.

These trends inform predictive use: Input employee data to flag high-risk individuals for targeted support.

Recommendations for Retention Strategies

Based on model insights and 2025 best practices, implement data-driven strategies to lower attrition to <10%, focusing on high-risk areas like onboarding and satisfaction.

1. Enhance Onboarding and Early-Tenure Support

- Develop mentorship programs for <5-year employees, with regular check-ins to reduce first-year exits by 20-30%. Target under-30s and single employees.

2. Career Development and Training

- Offer tailored programs in R&D/Sales, including internal promotions and skill-building, to address low satisfaction in roles like Lab Technicians. Track via KPIs like retention by role.

3. Improve Work-Life Balance and Benefits

- Introduce flexible models and wellness resources to counter overtime effects, monitored quarterly. Competitive compensation adjustments for high-risk groups.

4. Segmented Retention Initiatives

- Role-specific bonuses and recognition programs to build trust and engagement.

5. Promote Diversity and Inclusion

- Broaden recruitment for non-STEM roles and implement DEI training/mentorship to address male attrition dominance.

Implementation and Monitoring

Pilot in high-attrition departments, scaling with model predictions. Use quarterly dashboards to track metrics, leveraging AI for personalized strategies. Expected: 15-20% attrition drop, enhanced productivity.