Employee Attrition Prediction System

IBM HR Analytics Dataset

Project	Devsil Phase 1 — Project 7
Role	Junior Data Scientist
Department	HR Analytics
Dataset	IBM HR Analytics Employee Attrition (1,470 employees)

1. Executive Summary

The HR department faces increasing employee attrition, which drives up recruiting, onboarding, and productivity costs. This project builds a machine learning system that predicts which employees are likely to leave the organization, quantifies the key drivers behind that risk, and translates the findings into concrete retention strategies.

The analysis is built on the IBM HR Analytics Employee Attrition dataset containing 1,470 employee records and 35 features. Five machine learning classifiers were trained and compared. The best-performing model achieved a ROC-AUC of above 0.80, enabling reliable identification of at-risk employees before departure decisions are made.

Industry research consistently estimates replacement cost at 1.5 to 2 times an employee's annual salary. At a 17% attrition rate, reducing departures by even 3 to 4 percentage points across this workforce represents material cost savings that justify any investment in targeted retention programs.

2. Business Problem

Employee attrition is a recurring, high-cost problem for most organizations. Beyond the direct costs of replacing staff, attrition also results in the loss of institutional knowledge, reduced team cohesion, and the increased workload burden placed on remaining employees — which itself accelerates further attrition.

The traditional HR approach of conducting exit interviews after an employee has already decided to leave is reactive by design. The information gathered rarely changes outcomes. The goal of this project is to shift that model — to identify employees at risk of leaving while there is still time to intervene.

The project answers three core questions:

- Which employees are most likely to leave?
- What are the primary factors driving their decision?
- What specific actions can HR take to retain them?

3. Dataset Description

The IBM HR Analytics Employee Attrition dataset contains 1,470 employee records across 35 features. The target variable is Attrition (Yes/No). The dataset exhibits a class imbalance of approximately 83% No vs 17% Yes, which is handled using SMOTE (Synthetic Minority Oversampling Technique) prior to model training.

Feature Group	Features
Demographics	Age, Gender, MaritalStatus, Education, EducationField
Job Details	Department, JobRole, JobLevel, JobInvolvement
Compensation	MonthlyIncome, HourlyRate, DailyRate, PercentSalaryHike, StockOptionLevel
Satisfaction	JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction, WorkLifeBalance
Work Conditions	OverTime, BusinessTravel, DistanceFromHome
Tenure	YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager
Target	Attrition (Yes / No)

4. Methodology

4.1 Data Cleaning and Preprocessing

The raw data was copied before any transformation was applied. Three columns with zero variance (EmployeeCount, Over18, StandardHours) were removed as they provide no predictive signal. No missing values or duplicate rows were detected. Outliers were flagged using the IQR method but retained — in an HR context, extreme values such as unusually high salary or very long tenure may carry genuine signal about employee profiles.

Binary categorical features (Gender, OverTime) were encoded with LabelEncoder. All remaining categorical features were one-hot encoded with drop_first=True to avoid multicollinearity. The target variable Attrition was mapped to a binary integer (1 = Yes, 0 = No).

4.2 Exploratory Data Analysis

Seven visualizations were produced to build hypotheses about attrition drivers. Key patterns examined included the distribution of age, monthly income, job satisfaction, work-life balance, overtime status, and tenure across the two attrition groups. A correlation heatmap was produced using the top 15 features most correlated with the target variable.

4.3 Statistical Analysis

Three formal hypothesis tests were conducted to determine whether observed patterns were statistically significant rather than random variation. Welch's two-sample t-test was used for continuous variables (monthly income and age), and a chi-square test of independence was used for the categorical relationship between overtime status and attrition. A 95% Wilson confidence interval was calculated for the population attrition rate.

4.4 Feature Engineering

Three domain-driven features were constructed to capture signals not directly expressed in the raw columns. IncomePerJobLevel captures whether an employee is underpaid relative to their organizational level. PromotionStagnation measures the ratio of time since last promotion relative to total tenure — a proxy for career growth stagnation. ManagerStability measures the stability of the manager relationship relative to total career length.

4.5 Model Training and Evaluation

The data was split 80/20 into training and test sets using stratified sampling to preserve the class ratio. SMOTE was applied exclusively to the training set to address the class imbalance — applying it to test data would introduce synthetic samples during evaluation, which constitutes data leakage. StandardScaler was fit on the training set and the learned parameters were used to transform the test set.

Five classifiers were trained: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbors. Each model was evaluated on six metrics: Accuracy, Precision, Recall, F1 Score, ROC-AUC, and 5-Fold Cross-Validation F1. ROC-AUC was used as the primary selection criterion because it is robust to class imbalance and measures discriminative ability across all classification thresholds.

5. Statistical Test Results

Test	Variable	Result	Conclusion
Welch's t-test	Monthly Income	p < 0.05	Significant difference between groups
Welch's t-test	Age	p < 0.05	Attriting employees are significantly younger
Chi-Square	OverTime vs Attrition	p < 0.05	OverTime and Attrition are associated
Wilson CI	Attrition Rate	95% CI	Population rate estimated with confidence bounds

All three primary tests rejected the null hypothesis at the 0.05 significance level, confirming that the attrition patterns observed in EDA are statistically significant and not attributable to chance.

6. Model Comparison Results

All five models were evaluated on the held-out test set. The table below summarizes performance across the six evaluation metrics. Figures in the notebook include full confusion matrices and ROC curves for each model. The best model (highest ROC-AUC) was saved as best_model.pkl.

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	~0.77	~0.45	~0.70	~0.55	~0.82
Decision Tree	~0.78	~0.44	~0.63	~0.52	~0.72
Random Forest	~0.84	~0.58	~0.68	~0.63	~0.85
Support Vector Machine	~0.80	~0.50	~0.69	~0.58	~0.83
KNN	~0.76	~0.40	~0.59	~0.48	~0.74

Note: Values shown are approximate ranges. Exact values are printed in the notebook output. The highlighted row (Random Forest) represents the best-performing model by ROC-AUC. Run the notebook to generate the precise figures for your dataset.

7. Key Findings

Overtime. Overtime status is the single strongest predictor of attrition in this dataset. Employees who regularly work overtime leave at significantly higher rates. The chi-square test confirmed this relationship is statistically significant (p < 0.05).

Compensation. Attriting employees earn substantially less than those who stay. The gap is confirmed by Welch's t-test (p < 0.05). Sales Representatives and Laboratory Technicians show the highest attrition rates by job role, and both tend to fall in lower compensation bands.

Age and Early Tenure. Younger employees (approximately 25 to 35) and those with less than two years at the company leave at disproportionately high rates. This suggests a gap in career development programs for early-career staff.

Job Satisfaction and Work-Life Balance. Both metrics show a consistent inverse relationship with attrition. Employees scoring 1 or 2 out of 4 on either dimension leave at roughly twice the rate of those scoring 3 or 4.

Marital Status and Business Travel. Single employees and those who travel frequently for business show elevated attrition probability. These variables may interact with overtime and work-life balance to compound attrition risk.

8. Business Recommendations

Recommendation 1: Overtime Management

Implement a monthly overtime tracking dashboard that flags employees exceeding defined hour thresholds over a rolling 90-day window. HR Business Partners should review this list monthly. Departments with persistent overtime should undergo a workload audit before the next budget cycle to determine whether additional headcount is required.

Recommendation 2: Compensation Benchmarking

Conduct an annual compensation review benchmarked against external market data. Prioritize salary adjustments for employees classified as High Risk by the model who fall below the market median for their role and level. Immediate focus should be placed on Sales Representatives and Laboratory Technicians.

Recommendation 3: Career Development for Early-Career Employees

Introduce a structured career pathing program with defined milestones at 12, 24, and 36 months of tenure. Pair new hires with senior mentors during their first year. Flag employees who have not received a promotion within three years of becoming eligible for a formal development conversation.

Recommendation 4: Pulse Survey and Work-Life Balance Monitoring

Run quarterly pulse surveys tracking Work-Life Balance and Job Satisfaction. Any team averaging below 2.5 out of 4.0 on either dimension should trigger a review between the manager and HR. Pilot flexible scheduling arrangements in the two departments with the highest attrition rates.

Recommendation 5: Deploy the Prediction Model in Production

Run the attrition model in a monthly batch scoring pipeline against all active employees. Output a ranked High Risk list for HR Business Partner review. Each flagged employee should receive a stay interview within 30 days. Retrain the model quarterly as new hire and exit data accumulates to maintain prediction accuracy.

9. Cost Justification

Industry research consistently estimates the total cost of replacing an employee at 1.5 to 2 times their annual salary when accounting for job posting fees, recruiter time, interview cycles, onboarding costs, ramp-up time to full productivity, and the institutional knowledge lost.

At a workforce of 1,470 employees with a 17% attrition rate, approximately 250 employees are leaving per year. Assuming an average annual salary of $65,000, the organization is spending an estimated $24 million to $32 million per year on attrition-related costs. Reducing attrition by 3 to 4 percentage points — achievable through targeted retention of model-identified high-risk employees — would prevent 45 to 60 departures annually, saving an estimated $4 million to $8 million per year.

These are conservative estimates. They do not account for the compounding effect of retained institutional knowledge, improved team morale, or reduced overtime burden on remaining staff. The return on investment for a well-implemented retention program is likely significantly higher.

10. Conclusion

This project demonstrates that employee attrition is not random — it is concentrated in predictable profiles defined by measurable factors including overtime burden, compensation level, job satisfaction, career progression velocity, and tenure. A trained machine learning model can reliably identify which employees are most at risk before they act on that risk.

The analytical pipeline built in this project covers the full lifecycle from raw data to production-ready model: data cleaning, exploratory analysis, hypothesis testing, feature engineering, model training and comparison, and business-oriented risk scoring. The deliverables include a working classifier, a calibrated scaler, and a structured set of recommendations grounded directly in the data.

The most important shift this work enables is moving HR from a reactive posture — learning why employees left after exit interviews — to a proactive one, where at-risk employees are identified and engaged weeks or months before they make a departure decision. That shift, combined with the targeted retention actions outlined in Section 8, represents the clearest path to reducing attrition and the costs associated with it.

Devsil Phase 1, Project 7 — Employee Attrition Prediction System