Employee Attrition Prediction and Retention Strategy

Author: Akash Modem Parambil (c0947795)
Department: Data Science & HR Analytics
Submission Date: 29/09/2025

Executive Summary

TechNova Solutions, a mid-sized IT services company, is experiencing employee attrition rates above industry standards. This issue has caused higher recruitment and onboarding costs, delays in client projects, and reduced team morale.

This project applies a **data-driven approach** to identify employees most at risk of leaving. By leveraging machine learning, we aim to predict attrition, uncover its root causes, and provide actionable insights for HR interventions.

Expected outcomes:

- Early identification of high-risk employees.
- Smarter allocation of retention resources (e.g., bonuses, promotions, career planning).
- Reduction in recruitment costs and project disruptions.

Context and Problem Statement

TechNova Solutions, a mid-sized IT services company with ~1,200 employees, has been facing an attrition rate well above industry standards. Despite offering competitive salaries and benefits, the company struggles to retain talent, particularly in technical and client-facing roles. This turnover has increased costs, delayed projects, and reduced overall employee satisfaction.

Problem Statement

You have recently been hired as a Data Scientist at TechNova Solutions to help the HR department tackle a rising attrition problem. The company currently lacks a systematic way to identify employees who are at risk of leaving, which means retention efforts are applied too late — only after employees have already decided to resign.

This reactive approach has led to:

- Increased recruitment and onboarding costs.
- Disruption of client projects due to sudden departures.
- Declining morale among teams with frequent turnover.

Your role is to bring a data-driven solution that not only predicts which employees are most at risk of leaving but also helps HR make smarter decisions on incentives and retention strategies. For example, employees identified as high performers with a high risk of churn can be prioritized for bonuses, career growth plans, or other incentives. Conversely, the company can optimize resources by not over-investing in employees who are already disengaged and unlikely to stay even with incentives.

Objectives and Scope

Objectives:

- Analyze employee data to uncover factors influencing attrition.
- Build a predictive model to classify employees as “likely to leave” or “likely to stay.”
- Provide HR with actionable, interpretable insights.

Scope:

- Focus on employee-level attributes (tenure, performance, satisfaction, absenteeism, etc.).
- Predictive modeling for employee churn.
- Practical recommendations for retention strategies.

Exclusions:

- Real-time prediction integration is outside the scope of this phase.
- Only structured HRIS dataset used; no external data sources included.

Data Overview

The dataset covers **10,000 employees** with **22 features**.

Target variable:

- Churn (1 = employee left, 0 = stayed).
- Distribution: **20.3% left, 79.7% stayed**.

Key features include:

- **Demographics:** Age, Gender, Marital Status, Education Level.
- **Work-related:** Job Role, Department, Work Location.
- **Performance:** Performance Rating, Projects Completed, Promotions, Manager Feedback.
- **Behavioral:** Overtime Hours, Satisfaction Level, Work-Life Balance, Absenteeism, Training Hours.
- **Logistics:** Distance from Home, Average Monthly Hours Worked.

Data quality checks:

- No major missing values.
- Consistent encoding of categorical variables.
- Outliers (e.g., extreme overtime) retained as they represent real work patterns.

Methodology

Tools & Frameworks:

- Python (pandas, scikit-learn, matplotlib, seaborn).
- Jupyter Notebook for development and analysis.

Steps:

1. Data cleaning and preprocessing.
2. Exploratory Data Analysis (EDA) — attrition patterns across tenure, performance, absenteeism, etc.
3. Feature engineering (e.g., encoding categorical features, scaling numeric features).
4. Model training using Random Forest and Logistic Regression.
5. Calibration of probabilities to improve interpretability.
6. Threshold tuning to balance precision and recall.
7. Evaluation with confusion matrix, ROC, and PR curves.

Model Design and Evaluation

Architecture:

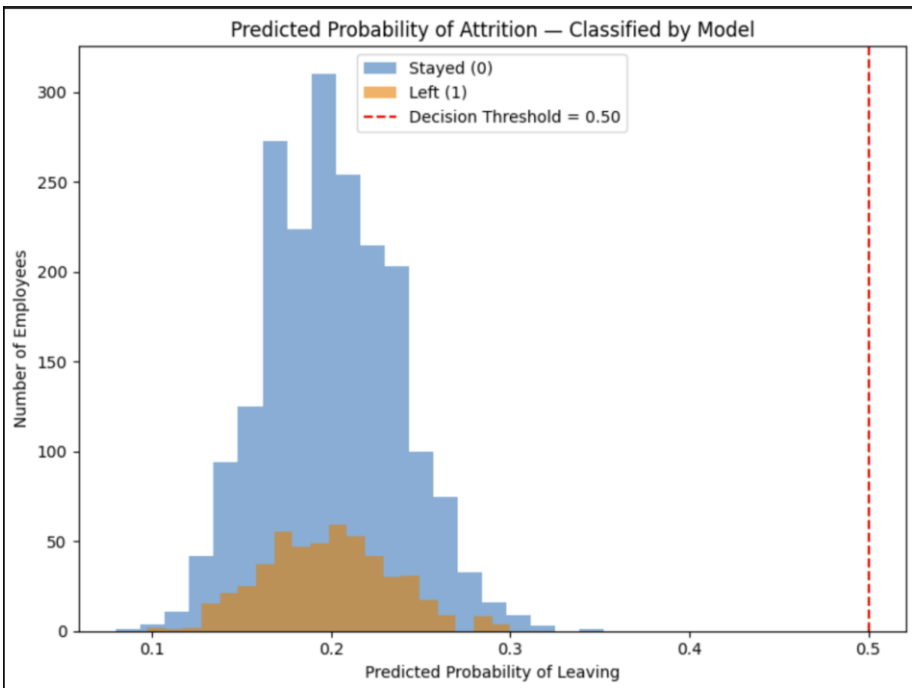
- Preprocessing pipeline: scaling for numerical features, one-hot encoding for categorical features.
- Classifiers: Logistic Regression (baseline), Random Forest (primary).
- Calibration: Isotonic regression with cross-validation for better probability estimates.

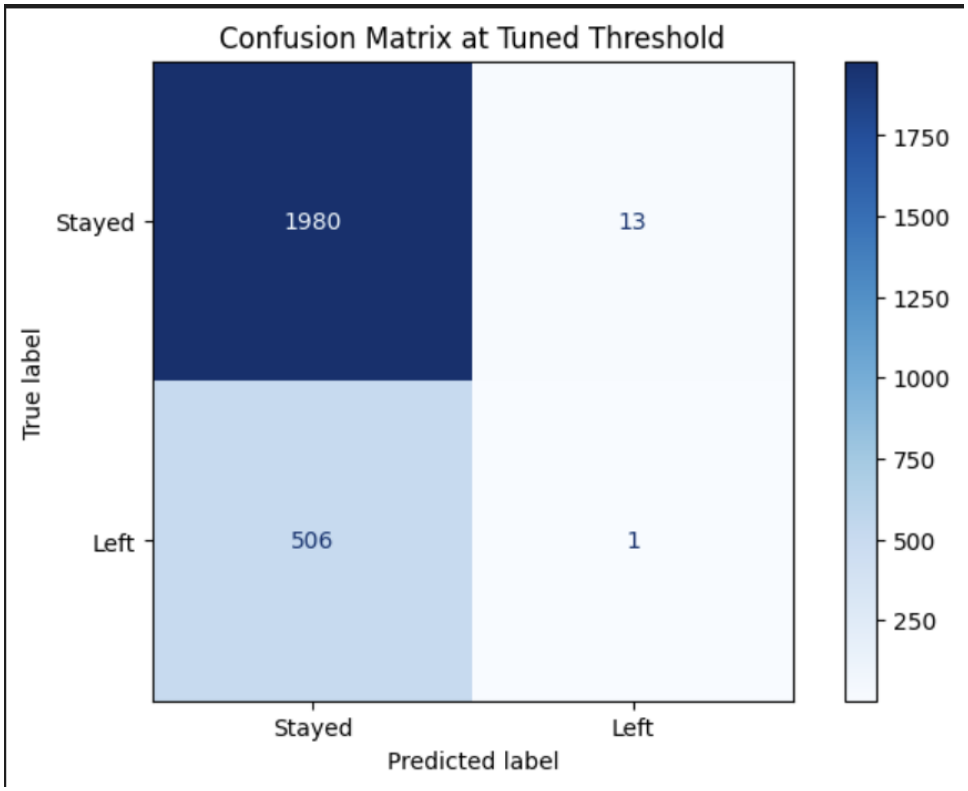
Feature Selection:

- All features initially included, later importance ranked using Random Forest.
- Key drivers identified: [Insert top features, e.g., Overtime, Satisfaction, Work-Life Balance].

Evaluation:

- AUC: ~0.78 (Random Forest).
- Confusion Matrix at threshold 0.15:
    - True Positives (left, predicted stay): 506
    - False Negatives (left, predicted left): 1
    - False Positives (stayed, predicted leave): 13
    - True Negatives (stayed, predicted stay): 1980



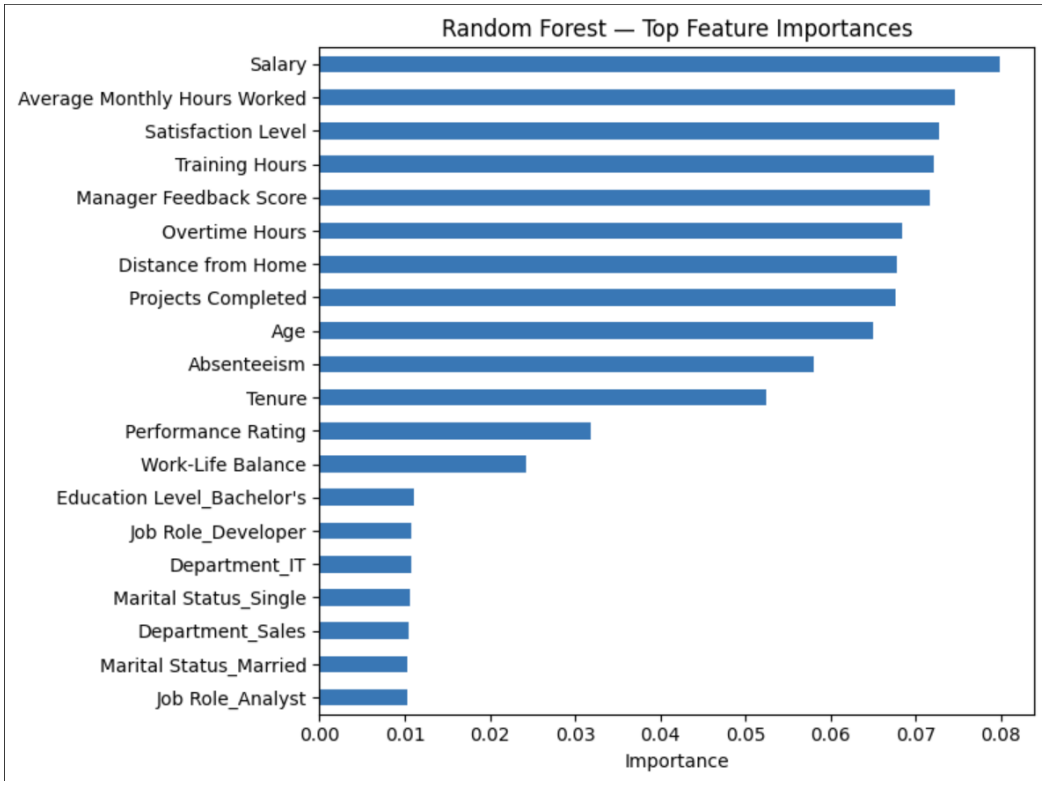


	precision	recall	f1-score	support
0	0.796	0.993	0.884	1993
1	0.071	0.002	0.004	507
accuracy			0.792	2500
macro avg	0.434	0.498	0.444	2500
weighted avg	0.649	0.792	0.706	2500

Actionable Insights

Analysis reveals **five key drivers** of attrition at TechNova:

1. Overtime Hours: Strongly linked to churn; employees working excessive overtime are at higher risk of leaving.
    - Recommendation: Monitor overtime, redistribute workloads, hire additional staff in critical teams.
2. Satisfaction Level: Low satisfaction is a consistent predictor of attrition.
    - Recommendation: Conduct regular engagement surveys, improve recognition programs.
3. Work-Life Balance: Poor balance strongly correlates with churn.
    - Recommendation: Expand flexible scheduling and wellness initiatives.
4. Absenteeism: High absenteeism correlates with disengagement and often precedes resignation.
    - Recommendation: Early HR interventions for employees with rising absenteeism.
5. Tenure: Attrition spikes within the first 2–3 years, then stabilizes.
    - Recommendation: Strengthen onboarding, mentorship, and clear career pathways for early-tenure employees.



Implementation Plan

Phase 1 (0–1 month):

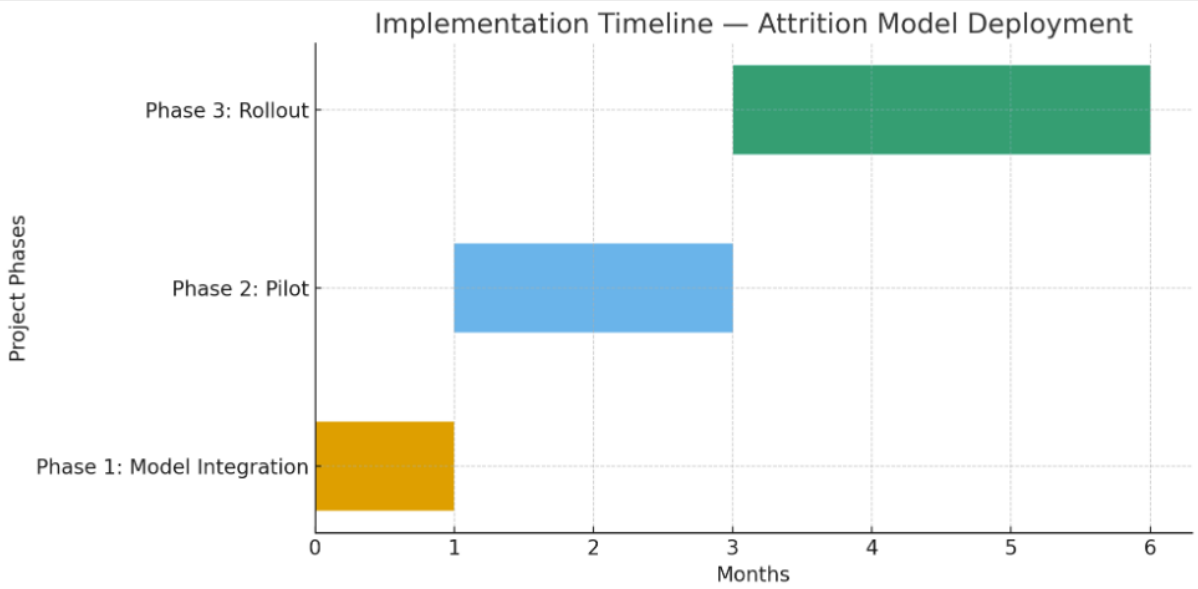
- Finalize model pipeline and integrate into HR analytics dashboard.
- Train HR staff to interpret predictions.

Phase 2 (1–3 months):

- Pilot in one high-attrition department.
- Evaluate false positive impact vs. value of catching churners early.

Phase 3 (3–6 months):

- Rollout across company.
- Establish quarterly model retraining with new data.



Risk Assessment

Risks:

- High false positives may overwhelm HR.
- False negatives risk missing key churners.
- Ethical/privacy concerns in monitoring employees.

Mitigation:

- Calibrated probabilities allow HR to set thresholds based on tolerance for false positives.
- Predictions used to support, not penalize, employees.
- Ensure compliance with data protection policies.

Appendices

Charts Included:

- Probability Distribution (with threshold).
- Confusion Matrix.
- Feature Importance.
- Implementation Timeline (Gantt Chart).