Lambton College

AML-3303: Software Tools and Emerging
Technologies for AI and ML

Technical Specification Document

Employee Attrition Prediction System

Author: Ryan Moses
Student ID: c0946432
Submission Date: 29/09/25

Contents

Executive Summary

TechNova Solutions is launching a data-driven initiative to address an ongoing employee retention issue. The project aims to build a predictive system that identifies employees at risk of leaving and provides HR with actionable insights to guide proactive retention strategies. By integrating data from HR systems, surveys, and project records, the system will support smarter, targeted interventions such as career development programs or workload adjustments.

The expected outcomes include reduced recruitment costs, greater stability in client projects, and higher employee satisfaction — positioning TechNova as a more resilient and attractive employer.

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. Measures like offering competitive salaries and benefits have not been effective in addressing this issue. Some challenges faced by the company due to this are

- Rising recruitment and onboarding costs.
- Delays in project delivery caused by sudden departures.
- Declining team morale due to constant turnover.

Currently, HR interventions are reactive, applied only after employees have already chosen to resign. Without a systematic approach to predict and address attrition risk, the company struggles to retain key talent in technical and client-facing roles. This project is intended to close that gap by enabling proactive, data-driven retention management.

Objectives and Scope

Objectives:

- Predict which employees are likely to leave.
- Identify main drivers of attrition.
- Support HR in designing proactive retention strategies.

Deliverables:

- A cleaned and well-documented dataset combining HR records, surveys, and workload data.
- An employee attrition prediction model with explainability features (e.g., SHAP values, feature importance).
- Risk scoring system with categories (Low, Medium, High).
- HR dashboard (Power BI) displaying churn risk, key drivers, and recommended actions.
- Deployment of a prediction service (batch + API) integrated with HR workflows.
- Documentation and training materials for HR staff.

Assumptions:

- Historical HR data is reliable and complete.
- HR will use predictions in combination with human judgment.

Data Overview

The dataset consists of employees information of TechNova Solutions which can be used for predicting employee attrition. Each row represents an individual employee, and the fields capture demographic details, work-related information, performance indicators, and satisfaction measures.

Key Details

- Number of Records: Each row corresponds to one employee.
- Target Variable: Churn (0 = Employee stays, 1 = Employee leaves).
- Feature Types:
  - Categorical: Employee ID, Gender, Job Role, Department, Work Location, Education Level, Marital Status.
  - Numeric: Age, Tenure, Salary, Training Hours, Overtime Hours, Absenteeism, Distance from Home, etc.
  - Ordinal/Rating Scales: Performance Rating (1–5), Work-Life Balance (1–4), Satisfaction Level (0–1), Manager Feedback Score (1–5).

Feature Categories

1. Demographics
   - Age, Gender, Education Level, Marital Status.
2. Job & Career Information
   - Tenure, Job Role, Department, Promotions, Work Location.
3. Compensation & Workload
   - Salary, Overtime Hours, Average Monthly Hours Worked.
4. Performance & Growth
   - Performance Rating, Projects Completed, Training Hours, Manager Feedback Score.
5. Engagement & Satisfaction
   - Satisfaction Level, Work-Life Balance, Absenteeism, Distance from Home.
6. Target Variable
   - Churn: Indicates whether the employee has left the company.

Methodology

Data Engineering

Data Collection

Aggregate employee data from HR records, surveys, and workload management systems.

Data Cleaning & Preparation

Remove duplicates, standardize formats, handle missing values through imputation, and filter outliers to ensure robust input for modeling.

- Handling Missing Values:
  - Impute numerical fields (e.g., Age, Salary) using mean/median.
  - Impute categorical fields (e.g., Department, Job Role) using most frequent value.
  - Drop Employee ID from modeling (identifier only).
- Categorical Encoding:
  - One-hot encoding for variables like Department, Job Role, Marital Status.
  - Ordinal encoding for fields with natural order (Education Level, Work-Life Balance, Performance Rating).
- Normalization / Scaling:
  - Apply standard scaling (z-score) or min-max scaling to continuous variables (e.g., Salary, Distance from Home, Training Hours).

Feature Engineering

Create and select relevant features capturing demographics, performance, engagement, and historical churn patterns.

- Tenure Buckets: Group Tenure into categories (e.g., 0–2 yrs, 3–5 yrs, 6+ yrs).
- Promotion Frequency: Promotions / Tenure (captures growth pace).
- Performance Trend: Rolling average or slope of performance ratings if multi-year data available.
- Overtime Intensity: Overtime Hours ÷ Average Monthly Hours.
- Engagement Index: Weighted score combining Satisfaction, Work-Life Balance, and Training Hours.

Exploratory Data Analysis (EDA)

- Visualize trends in attrition by department, tenure, and engagement levels.
- Identify initial key drivers using correlation analysis and summary statistics.

Predictive Modeling

- Model Selection: Prototype multiple algorithms (Logistic Regression, Random Forests, Gradient Boosted Trees, and Neural Networks) using Python-based frameworks (scikit-learn, XGBoost, TensorFlow).
- Model Training: Split cleansed data into train, validation, and test sets (e.g., 60-20-20). Apply cross-validation to fine-tune hyperparameters and mitigate overfitting.
- Model Interpretability: Use explainability techniques (e.g., SHAP values, feature importance) so HR can understand and trust risk predictions.

Model Evaluation

- Use key metrics—Recall (to minimize false negatives), F1 Score, ROC-AUC—to assess and select the best model.
- Confirm stability of results and fairness across demographic subgroups.

Model Design and Evaluation

Candidate Algorithms

- Logistic Regression: Baseline interpretable model.
- Random Forest: Handles non-linear relationships, good with mixed feature types.
- XGBoost / LightGBM: Strong gradient boosting methods, often best for tabular churn data.
- Neural Networks (MLP): May capture complex interactions but requires tuning and larger data.

Evaluation Metrics

- Recall (Sensitivity) for “At Risk” employees → prioritize minimizing false negatives.
- F1-score → balances precision and recall.
- ROC-AUC → measures overall separability of churn vs stay.
- Precision-Recall AUC → more useful if attrition is rare (class imbalance).

Model Interpretability

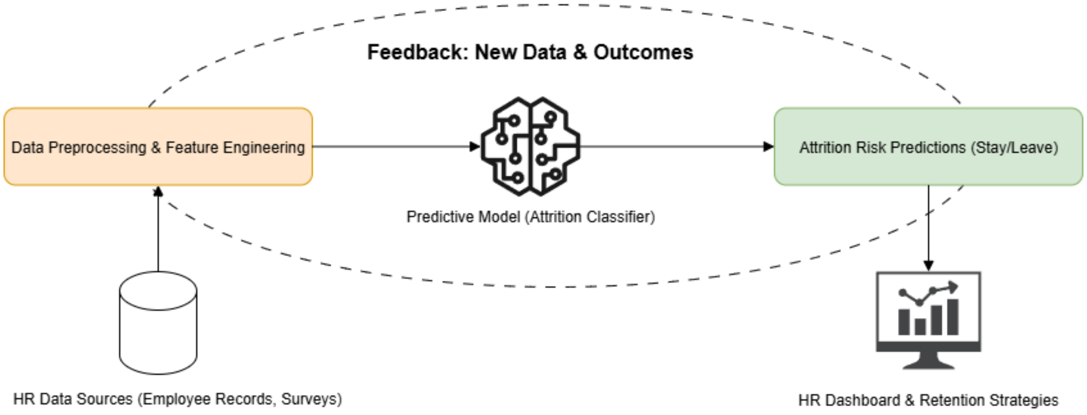
- Feature Importance: From tree-based models (Random Forest, XGBoost).
- SHAP values: Explain individual predictions (why an employee is flagged at risk).

Actionable Insights

1. Prediction Score (Probability of Attrition)
    - The predictive model outputs a probability (e.g., 0.78) representing how likely an employee is to leave.
    - This value flows into the HR Dashboard.
2. Risk Categorization (Low, Medium, High)
    - The probability is transformed into clear categories using predefined thresholds.
    - Example:
        - Low: < 0.3
        - Medium: 0.3–0.6
        - High: > 0.6
    - These categories make results more interpretable for HR managers.
3. Recommended Interventions
    - Alongside predictions, the system suggests data-driven retention strategies such as:
        - Career growth opportunities
        - Salary adjustment
        - Mentorship programs
        - Engagement/recognition initiatives
    - These recommendations are displayed on the HR Dashboard to support decision-making.

Implementation Plan

- System Architecture:


Feedback: New Data & Outcomes
Data Preprocessing & Feature Engineering
Predictive Model (Attrition Classifier)
Attrition Risk Predictions (Stay/Leave)
HR Data Sources (Employee Records, Surveys)
HR Dashboard & Retention Strategies

- Deployment Steps:
  1. Build ETL pipelines (Azure Data Factory).
  2. Train models (Azure ML / Databricks).
  3. Deploy scoring service (AKS endpoint).
  4. Integrate predictions with Power BI dashboards.
  5. Establish feedback loop for retraining.
- Timeline:
  - Data Preparation – 2 weeks
  - Modeling & Validation – 4 weeks
  - Dashboard & Deployment – 3 weeks
  - HR Training & Rollout – 1 week

Risk Assessment

1. Biased Predictions (e.g., gender, age)
    - Risk: Model may learn unwanted patterns from biased historical HR data.
    - Mitigation:
        - Perform fairness audits on model outputs.
        - Apply bias mitigation techniques (re-sampling, re-weighting, fairness constraints).
        - Exclude sensitive attributes from training features.
2. Overfitting due to Limited Data
    - Risk: Model performs well on training data but poorly on new employees.
    - Mitigation:
        - Use cross-validation and regularization.
        - Collect and integrate more diverse HR data over time.
        - Apply early stopping during training.
3. False Positives (incorrectly flagging employees as high attrition risk)
    - Risk: Wasted retention resources, reduced trust in the system.
    - Mitigation:
        - Set appropriate decision thresholds to balance precision and recall.
        - Provide confidence scores with predictions.
        - Use human-in-the-loop review before major HR interventions.