Proactive Attrition Risk System: Project Documentation & Technical Report

Section	Detail
Project Name	Proactive Attrition Risk System
Supervised By	Dr. Mahmoud Talaat
Prepared By	Ahmed Fekry Mohamed (Team Leader), Yousif Moaz ELbadry, Mazen Kamal Morsy, Shaimaa Ahmed Fouad, Youssef Ahmed Fouad
Date	November 2025

1. Executive Summary

Employee attrition represents a significant financial and operational risk, driving up recruitment costs and eroding institutional knowledge. This project successfully developed and deployed a **Proactive Attrition Risk System**, a machine learning-based solution designed to transform Human Resources strategy from reactive to proactive.

The system utilizes a novel **Two-Stage Predictive Engine** that combines supervised learning (XGBoost) with unsupervised segmentation (PCA + Clustering). The output not only accurately forecasts **who** is likely to leave but, critically, categorizes them into actionable groups (e.g., "Career Stagnation Group," "Logistical Burden Group"), enabling highly targeted and cost-effective retention strategies. The final model achieved a high Recall of over 80%, meeting the core objective of capturing the majority of potential employee exits.

2. Project Objectives and Scope

The core purpose of this project is to counteract the financial and operational risks associated with high talent turnover.

2.1 Core Objectives

1. **Discover Causes:** Identify the key underlying drivers of attrition (e.g., low pay, lack of growth, work-life imbalance).
2. **Forecast Risks:** Accurately predict the probability of an employee leaving within a defined future period using advanced ML models.
3. **Empower Action:** Provide HR managers with an actionable dashboard and segmented employee groups to facilitate timely, targeted interventions.

2.2 Key Performance Indicators (KPIs)

The project measured success based on both technical rigor and business impact.

KPI Category	Metric	Target / Result
Business Impact	Model Recall	**> 80% (Achieved)**
Business Impact	Cost Reduction	Target: 15% reduction in annual turnover costs
Model Performance	F1-Score	**77%** (Best balance of Precision and Recall)
Data Quality	Data Accuracy after Preprocessing	99%
Technical Design	Prediction Speed (Latency)	~1 millisecond

3. System Methodology and Design

The system is built on a robust pipeline that processes raw HR data through specialized machine learning stages.

3.1 Data Pipeline and Preprocessing

Step	Description	Rationale
Data Source	Synthetic dataset based on the IBM HR Attrition & Performance schema, scaled up to **~250,000 samples** to simulate realistic organizational dynamics.	Validation of the engine logic and testing segmentation strategies without using sensitive employee data.
Encoding	Qualitative variables (e.g., `BusinessTravel`, `Department`) converted into numerical vectors.	Ensure compatibility with all machine learning algorithms.
Feature Scaling	**Standardization** applied to features with vast ranges (`Monthly Income`, `Age`).	Prevent bias introduced by disparate feature scales and improve model convergence speed.
Class Imbalance	**SMOTE (Synthetic Minority Over-sampling Technique)** applied to the training set. (Attrition is a rare event, ~13%).	Prevents the model from ignoring the high-risk minority class, which is crucial for achieving high Recall.

3.2 The Two-Stage Predictive Engine

The system's innovation lies in segmenting the high-risk population for intervention, not just identifying them.

Stage 1: Supervised Prediction (XGBoost)

The objective of this stage is to answer the question, **"Who is likely to leave?"**

- **Algorithm Selection:** XGBoost (Extreme Gradient Boosting) was selected after rigorous cross-validation and benchmarking against Random Forest, Logistic Regression, and K-NN.
- **Performance Rationale:** While Logistic Regression achieved a decent ROC AUC due to SMOTE, XGBoost offered a significantly superior **Precision (76%)** and **F1-Score (77%)** by finding complex, non-linear relationships, drastically reducing false alarms compared to simpler models.
- **Output:** A list of employees ranked by their **Attrition Risk Score** (probability).

Stage 2: Unsupervised Segmentation (PCA & Clustering)

The objective of this stage is to answer the question, **"Why are they leaving?"**

1. **Dimensionality Reduction (PCA):** Principal Component Analysis was applied to the features of the high-risk cohort.
   - **Challenge:** The total explained variance was **62.28%** for the three selected components.
   - **Solution:** We prioritized **interpretability** over total variance. We successfully mapped the three components to the strongest original attrition-driving features (`Monthly Income` , `Job Level` , `Years at Company`).
2. **Clustering (K-Means):** The reduced feature space (the 3 Principal Components) was used to group employees into distinct, meaningful clusters.
3. **Actionable Segmentation:** The resulting clusters are translated into actionable HR categories for targeted intervention strategies:
   - **Example Cluster:** Employees characterized by low recent promotions, long tenure, and flat salary trajectory (The "Career Stagnation" Group).

4. UI/UX Design and Deployment

The system is deployed via a secure, web-based dashboard designed for clear, actionable decision-making.

Interface Component	Stakeholder	Purpose
Main Dashboard	HR Managers	Provides a company-wide risk overview and high-level attrition trends.
Employee Profile View	Team Leads / HR	Shows individual risk scores, the **top 3 contributing risk factors**, and suggested intervention actions.
Cluster View	HR Strategists	Visualizes the 3D PCA clusters to identify systemic issues and validate segmentation effectiveness.
Alerts	Managers	Automatic notifications when an employee's risk score crosses a predefined threshold.

4.1 Tools and Technologies

Category	Tools / Languages
Programming	Python (Pandas, NumPy, Scikit-learn, XGBoost)
Data Processing	`imblearn` (for SMOTE), `StandardScaler`
Development	Google Colab, ANACONDA
Deployment	Docker, Cloud Services (for MLOps pipeline)

5. Conclusion

The Proactive Attrition Risk System is a fully operational, high-value asset that delivers on its promise of shifting the company's talent management strategy. By moving from a reactive position—only understanding why people *left*—to a **proactive system** that forecasts *who* will leave and *why*, the organization is empowered to:

- **Preserve Key Talent:** Intervene with specific strategies tailored to the individual's root cause of risk.
- **Reduce Financial Strain:** Significantly cut down the recurrent costs associated with recruitment and training new employees.

The system is a direct tool to protect and maximize human capital investment.