PROJECT REPORT — Workforce Analytics & Attrition Optimization

Driving proactive employee retention through explainable machine learning and business intelligence

Prepared by: Shreya Deshpande
Target Role: Data Scientist
Toolset: Python, Pandas, NumPy, Scikit-learn, SQL, Power BI
Dataset Source: Kaggle — *HR Analytics Employee Attrition Dataset*
https://www.kaggle.com/datasets/HRAnalyticRepository/employee-attrition-data

Executive Summary

Employee attrition represents a critical organizational risk, impacting productivity, institutional knowledge, hiring costs, and long-term business continuity. This project addresses attrition as a **predictive and preventable business problem**, rather than a reactive HR metric.

Using a combination of **machine learning models** and **Power BI dashboards**, this solution:

- Identifies employees at elevated attrition risk
- Explains *why* attrition risk exists using interpretable ML
- Translates predictive outputs into **actionable decision support** for HR and leadership

The final deliverable integrates:

- Robust data preparation and feature engineering
- Predictive modeling using Logistic Regression and Random Forest
- Explainability through feature importance and coefficients
- A structured Power BI dashboard designed for stakeholder consumption

This project demonstrates end-to-end ownership of the data science lifecycle — from raw data to executive-ready insights.

Dataset Overview

The dataset contains historical workforce records capturing employee demographics, job attributes, tenure, organizational placement, and attrition outcomes.

Dataset Characteristics

- **Total records:** 49,648 employees
- **Attrition rate:** 2.99%
- **Target variable:** `attrition_flag` (1 = exited, 0 = retained)
- **Time span:** Multi-year employee history

Attribute Categories

Category	Example Attributes
Demographics	Age, Gender
Employment Details	Job Title, Department
Organizational Placement	City, Store, Business Unit
Tenure & Experience	Length of Service, Year
Outcome	Attrition Flag

The dataset exhibits **strong class imbalance**, with attrition events representing a small but high-impact minority — a realistic and critical modeling challenge in HR analytics.

Data Preparation & Feature Engineering

Significant preprocessing was performed to ensure analytical integrity and modeling reliability.

Key Preparation Steps

- **Data validation:** Row counts verified post-load (49,648 records)
- **Attrition flag encoding:** Binary classification target created
- **Temporal engineering:** Year extracted for trend analysis
- **Tenure metrics:** Length of service standardized and validated
- **Categorical encoding:** One-hot encoding applied to job title, department, city, and gender
- **Null handling:** No missing values detected in critical fields
- **Data consistency checks:** Workforce totals reconciled across years

Result

A fully structured, ML-ready dataset aligned for:

- Predictive modeling
- Explainability analysis
- Business intelligence visualization

Exploratory Workforce Analysis

Before modeling, exploratory analysis was conducted to understand baseline workforce dynamics.

Key Observations

- Workforce size increased steadily before stabilizing
- Attrition events remained a small percentage but consistent over time
- Certain job roles dominate workforce composition, indicating role-specific exposure
- Average tenure varies significantly across years, suggesting structural retention differences

Referenced Visuals

- *Workforce Headcount by Year*
- *Active vs Terminated Employees Trend*
- *Average Tenure by Year*
- *Employee Distribution by Job Title*

These insights informed both feature selection and modeling strategy.

Predictive Modeling Approach

Two classification models were developed and evaluated to predict attrition probability.

Model 1 — Logistic Regression

- ROC-AUC: **0.882**
- Strong baseline interpretability
- High recall for attrition cases
- Lower precision due to class imbalance

Model 2 — Random Forest

- ROC-AUC: **0.937**
- Superior discrimination power
- Improved balance between precision and recall
- Robust handling of non-linear relationships

Model Selection Rationale

Random Forest was selected as the primary model because it:

- Significantly outperformed Logistic Regression
- Provided stable and consistent predictions
- Allowed feature-level explainability via importance scores

This ensured predictive strength **without sacrificing transparency**.

Model Evaluation & Performance

Random Forest Performance Summary

- **Accuracy:** 98%
- **Attrition Recall:** 73%
- **Attrition Precision:** 66%
- **Confusion Matrix:** Demonstrates strong true positive identification with controlled false positives

Given the business objective — *early identification of at-risk employees* — recall was prioritized over raw accuracy.

Explainability — Key Attrition Drivers

To avoid black-box decision-making, explainability was treated as a **non-negotiable requirement**.

Explainability Methodology

- Random Forest feature importance
- Logistic Regression coefficient magnitude (directional validation)
- Cross-model consistency checks

Top Drivers Increasing Attrition Risk

- Age
- Specific city locations
- Job roles classified as “Other” (Miscalleneous roles)
- Certain calendar years (organizational effects)

Top Drivers Reducing Attrition Risk

- Longer length of service
- Executive and Investment departments
- Stable geographic placements

Referenced Visual

- *Model Explainability — Key Attrition Drivers*

Business Interpretation

Employees with **shorter tenure**, **role instability**, and placement in **specific departments or locations** show consistently higher attrition risk.
These drivers appear across both models, confirming **structural patterns rather than model noise**.

This section establishes responsible ML practices and elevates the project beyond standard predictive dashboards.

Power BI Dashboard — Design & Documentation

The Power BI dashboard translates model outputs into **decision-ready insights**, designed for HR leadership and business stakeholders.

Section A — Workforce Overview

Purpose: Establish organizational context and baseline trends.

Key Visuals

- Workforce Headcount by Year
- Active vs Terminated Employees
- Average Tenure Trend

Business Value
Provides macro-level understanding before risk segmentation.

Section B — Attrition Risk Distribution

Purpose: Understand how risk is distributed across the workforce.

Key Visuals

- Predicted Attrition Probability Distribution
- Risk Band Segmentation

Business Value
Enables early intervention planning by identifying risk concentration.

Section C — Risk by Business Dimension

Purpose: Reveal structural risk patterns.

Key Visuals

- Average Predicted Risk by Department *(with tooltip)*
- Average Predicted Risk by Job Title *(with tooltip)*
- Average Predicted Risk by Business Unit
- Average Predicted Risk by City

Tooltip Design
Tooltips provide:

- Risk band breakdown
- Average tenure context
- Employee count context

This allows deep analysis **without overcrowding the dashboard**.

Section D — Model Explainability Panel

Purpose: Build trust in predictive outputs.

Key Visuals

- Top Attrition Drivers (Feature Importance)
- Business Interpretation Text Panel

Business Value
Demonstrates ethical, transparent AI usage suitable for HR decision-making.

Section E — Actionable Employee Risk Table

Purpose: Enable intervention.

Key Visual

- Employee Risk Table (sorted by attrition probability)

Design Rules

- Employee ID appears only here
- Conditional formatting highlights high-risk cases
- Supports targeted retention actions

Key Performance Indicators (KPIs)

Category	KPI
Workforce	Attrition Rate
Risk	High-Risk Employee Percentage
Tenure	Average Length of Service
Model	ROC-AUC Score
Explainability	Top Driver Stability

Value Delivered

This project delivers tangible organizational value:

✔ Predictive identification of attrition risk
✔ Transparent, explainable ML outputs
✔ HR-friendly dashboards for strategic planning
✔ Reduced reliance on reactive retention strategies
✔ Strong alignment between data science and business objectives

---

Conclusion

This project demonstrates advanced competency across:

- Workforce analytics
- Predictive modeling
- Explainable machine learning
- Business intelligence storytelling

By integrating **ML explainability with Power BI decision support**, the solution bridges the gap between data science and real-world organizational impact.