Capstone Project Proposal

Name: Divine Obadeyin
Program: AltSchool of Data Science (Baraka)

Predicting Employee Attrition Using Workforce Data

Problem Statement

Employee attrition is a common challenge in many organisations, affecting productivity, planning, and team stability. Companies collect a lot of employee information — job role, income, workload, satisfaction scores, and years at the company — but that data is not always used to anticipate who might leave. For this capstone, I decided to predict employee attrition using structured HR data. I'm treating it as a classification problem where the model learns patterns that separate employees who left from those who stayed.

Instead of examining one factor at a time, the dataset makes it possible to look at how several factors work together and how they relate to attrition risk. The goal is not to build a perfect HR decision system, but to apply a clean, reproducible machine learning workflow and evaluate how well standard classification models perform on real workforce data.

Why This Problem Matters

I chose this problem because employee turnover has a direct business cost. Replacing staff takes time, money, and training effort. Frequent exits can also affect morale and performance across teams. If risk patterns can be identified early, organisations can respond with better retention strategies — such as role adjustments, workload balancing, or incentive programs. This is also a good evaluation case because prediction errors are not equal. Missing a high-risk employee matters more than raising a false alarm, which makes metric choice important.

Project Objectives

The main objectives of this project are to:

- Obtain a publicly available employee attrition dataset and review its structure
- Clean and prepare the dataset for modelling
- Explore which employee and job factors relate most to attrition
- Train a baseline classification model
- Train at least one comparison model
- Evaluate models using multiple classification metrics
- Review model errors and feature importance
- Present results with clear visuals and a written explanation

---

Research Questions

This project will be guided by the following questions:

- Which employee characteristics are most associated with attrition?
- Can a classification model predict attrition risk with reasonable reliability?
- How do different models compare on this dataset?
- What types of prediction errors appear most often?
- What practical retention insights can be drawn from the results?

Proposed Dataset Source

I will use a publicly available HR Employee Attrition dataset from a recognised open data platform, **Kaggle.com**. The dataset contains employee-level records covering demographic details, job information, compensation, and satisfaction indicators. Typical variables include age, department, job role, monthly income, years at company, overtime status, and job satisfaction scores. The target variable is binary and indicates whether the employee left or stayed.

This makes it directly suitable for a classification task.

Link to the dataset:
https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

---

Success Metrics

Model performance will not be judged by accuracy alone. Attrition datasets are usually imbalanced, so additional metrics are necessary.

I will use:

- Recall — how many actual attrition cases are correctly identified
- Precision — how reliable attrition predictions are
- F1 score — balance between precision and recall
- Confusion matrix — to understand error types
- ROC–AUC — overall class separation ability

A model will be considered successful if it improves on a simple baseline and shows balanced performance across these measures.

Method Overview

The project will follow a standard end-to-end workflow:

- Data loading and inspection
- Data cleaning and consistency checks
- Handling missing or irrelevant fields
- Encoding categorical variables
- Feature preparation and scaling where needed
- Train/test split
- Baseline logistic regression model
- Tree-based comparison model
- Model evaluation and comparison
- Feature importance and error review

Constraints and Assumptions

The dataset comes from a public source, so its size and feature coverage are fixed. It may not include every factor that influences why employees leave, such as management quality or personal reasons. Because of time limits, model tuning will be practical rather than exhaustive. I also assume the recorded labels are mostly correct and consistently entered.

Model outputs will be treated as analytical insights, not final HR decisions.

Risks and Limitations

- Possible class imbalance affecting model learning
- Missing contextual factors not captured in the dataset
- Relationships observed will be correlational, not causal
- Results may not generalise to every organisation

These limits will be clearly stated when interpreting results.

Expected Outcome

The expected outcome is a reproducible classification workflow and a small set of evaluated models for predicting employee attrition. I expect to identify the most influential factors linked to attrition and compare how different models behave on the same data.

The final report will present the results with clear visuals and plain-language explanations so a non-technical reader can still follow the conclusions.