Predictive HR Analytics using AWS SageMaker

- **Title**: Predictive HR Analytics using AWS SageMaker
- **Course**: AAI-540 – Machine Learning Design
- **Subtitle**: Employee Attrition Prediction Model
- **Team Info**: Group # Individual Project 3
- **Authors**: Sanjay Kumar, Syed Ahmed Ali
- **Business Entity**: Tatas India Ltd.
- **Date**: February 22nd, 2026

Project Overview & Objectives

Employee Attrition Challenge

- Employee attrition leads to: High recruitment and onboarding costs, Loss of organizational knowledge, Reduced team productivity and morale
- Organizations lack **proactive visibility** into attrition risk
- HR decisions are often reactive instead of data-driven

Primary Objective: Develop and operationalize a production-grade ML model to predict voluntary employee attrition (Attrition = 1: "Left" vs. Attrition = 0: "Stayed").

Business Goal: Serve as a proactive decision-support tool for HR to identify high-risk employees before they resign.

Impact: Reduce voluntary turnover rates and associated replacement costs through targeted retention interventions.

Project Background & Motivation

The Problem: High attrition leads to recruitment costs, loss of institutional knowledge, and team disruption.

The Solution: A supervised binary classification model that learns historical patterns to forecast future risk.

Focus: Leveraging demographic, compensation, and performance factors for data-informed interventions.

Business Value

- Early identification of high-risk employees
- Enables targeted retention strategies:
- Career development
- Compensation review
- Workload balancing
- Recognition programs

Project Technical Scope

In Scope

- End-to-end ML lifecycle on AWS SageMaker
- Data ingestion, preprocessing, feature engineering
- XGBoost model training and evaluation
- Model registry and real-time deployment
- Monitoring and CI/CD design

Out of Scope

- Real employee data (only synthetic dataset)
- Time-to-attrition prediction
- Advanced fairness mitigation
- Front-end dashboards

Success Criteria

Technical Metrics

- ROC-AUC ≥ 0.65
- Recall ≥ 0.60 for attrition class
- F1-score ≥ 0.60

Operational Metrics

- Real-time inference latency < 500 ms
- Stable SageMaker endpoint
- Model registered and approved

Business & Ethics

- Interpretable feature importance
- Awareness of proxy bias risks
- No PII or PHI usage

Dataset Overview

Data Source

- Synthetic Employee Attrition Dataset (Kaggle)
- ~74,500 records
- License: CC0 (Public Domain)

Target Variable

- Attrition (Stayed / Left)

Key Feature Categories

- Demographics
- Compensation & tenure
- Job characteristics
- Satisfaction & perception
- Performance & growth

Data Engineering & Storage

- **Storage Architecture:** Amazon S3 serves as the single source of truth for raw data, processed sets, and model artifacts.
  +2
- **Data Lineage:** Raw CSVs stored under `raw-data/`, while SageMaker-compatible headerless CSVs are stored under `processed/`.
- **Tooling:** Integrated with SageMaker notebooks and S3-centric workflows for scalability.

Preprocessing Steps

- Missing value imputation - Median imputation for numerical features and mode imputation for categorical data.
- Ordinal, binary, and one-hot encoding
- Feature scaling using StandardScaler
- Stratified train/validation/test split - 70% Train, 15% Validation, and 15% Test split preserving class ratios.

Feature Engineering Strategy

Key Decisions

- Dropped Employee ID (non-informative)
- Ordinal encoding for ranked categories - Manual mapping for ranked categories (e.g., Work-Life Balance: Poor=1 to Excellent=4).
- Binary encoding for Yes/No features,
- One-hot encoding for nominal variables - Applied to nominal variables like Job Role and Marital Status.
- Dimensionality: Expanded from 23 original predictors to ~30–35 final columns.

Reasoning

- XGBoost handles non-linear interactions effectively
- Avoids overfitting on synthetic data

Model Selection & Training

Algorithm: XGBoost (Extreme Gradient Boosting).

Why XGBoost? Strong performance on tabular data, native handling of missing values, and efficient execution in SageMaker.

Infrastructure: Managed SageMaker training jobs using `ml.m5.large` instances.

Parameters: `binary:logistic` objective with 200 rounds and a max depth of 5.

Key Hyperparameters

- Objective: `binary:logistic`
- Eval Metric: `auc`
- Max Depth: 5
- Learning Rate (eta): 0.1
- Boosting Rounds: 200

Model Evaluation Results

Test Set Performance

- ROC-AUC ≈ **65.2%**
- Accuracy ≈ **65.4%**
- Precision ≈ **64.2%**
- Recall ≈ **61.4%**
- F1-score ≈ **62.8%**

Conclusion

- Model outperforms random and baseline classifiers
- Balanced trade-off between false positives and false negatives
- **Outcome:** The model successfully met the success criteria gate (ROC-AUC ≥ 0.65).

Model Deployment

Deployment Strategy

- Real-time SageMaker endpoint
- CSV input → JSON output
- Endpoint name: `employee-attrition-xgb-endpoint`
- **Latency:** Optimized for scoring single employee records in under 500 ms.
- **Model Registry:** Registered with "Approved" status, ensuring version control and auditability.

Why Real-Time?

- Interactive HR decision support
- Fast scoring during retention planning
- Easy REST API integration

Monitoring & CI/CD

Endpoint Health: CloudWatch metrics for latency, invocation rates, and error tracking.

Performance Monitoring: Periodic batch scoring to detect drift in AUC or Recall.

Automation: Planned CI/CD hooks for automated regression testing and model promotion

Model Monitoring

- Prediction score drift
- Performance degradation checks
- Planned integration with SageMaker Model Monitor

Source Control

- GitHub for notebooks and scripts

Risks & Future Enhancements

Ethical & Bias Risks

- Proxy attributes: age, gender, marital status
- Risk of disparate impact
- **Data Privacy:** Used synthetic data to eliminate PII/PHI risks.
- **Limitations:** Synthetic data may miss nuanced real-world regional or economic cycles.

Future Enhancements

- Hyperparameter tuning - Implement Bayesian Hyperparameter Optimization for better tuning.
- Fairness analysis & SHAP explainability - Integrate SHAP-based services to explain the "why" behind individual risk scores.
- Automated retraining pipelines - Build a full SageMaker Pipeline for automated retraining triggered by data drift.
- **Ensembling:** Experiment with stacking XGBoost with CatBoost or TabNet for higher accuracy.
- HR dashboards (QuickSight / Streamlit)

THANK
YOU