**AAI-540 ML Design Document**

**Soccer Match Outcome Prediction**

University of San Diego

Applied Artificial Intelligence

1. **Team Information**

Project Team Group #: 5

Authors: Kay Michnicki, Matt Thompson

Business Name: Predicting Football Match Outcomes

Publication Date: October 16, 2025

1. **Team Workflows**

GitHub Project Link: https://github.com/KayMichnicki/AAI540-Final-Project-Group5

Asana Board Link: https://app.asana.com/1/1211354556818329/project/1211354557568622/list/1211354557568635

Team Tracker Link: https://docs.google.com/document/d/1uxLwmREhOPBZhq0v-FwusmseajjbuVDiipdwhSPEMuY/edit?tab=t.0#heading=h.okf28hah9325

1. **Project Scope**
2. **Project Background**

The objective of this project is to build a machine learning model that predicts the outcomes of international soccer (football) matches. Specifically, the model forecasts whether the home team wins, the away team wins, or if the match ends in a draw. The system leverages over 150 years of match data to detect patterns that influence outcomes, such as team performance, ranking differentials, and home advantage.

This predictive model can be used for sports analytics, pre-match forecasting, and strategic planning for analysts and fans.

1. **Technical Background**

This is a multi-class classification problem with three classes: Home Win, Draw, and Away Win. Data Source: International Football Results dataset (Kaggle, 1872–2025). Evaluation Metrics: Accuracy, precision, recall, macro-F1 score, and confusion matrix.

Data Preparation includes handling missing values, cleaning team names, encoding categorical variables (e.g., teams, competition type), and creating engineered features such as ELO ratings, recent performance, and head-to-head stats. Exploration includes trend analysis over time, distribution of outcomes, and correlations between features and match results. Hypothesized Key Features: ELO rating differences, home advantage, recent match form, tournament type. Model Choice: Logistic Regression as a baseline, and XGBoost/LightGBM for improved performance and robustness.

1. **Goals vs Non-Goals**

Goals

• Accurately predict match outcomes (Home/Draw/Away).

• Create informative, engineered features.

• Demonstrate a robust multi-class classification pipeline.

• Ensure reproducibility with Git version control and SageMaker pipelines.

Non-Goals

• Real-time prediction at scale.

• Player-level performance analysis.

• Financial or betting recommendation system.

• Perfect accuracy — focus on explainable performance.

1. **Solution Overview**

We designed a time-aware classification pipeline that uses only pre-match information to avoid leakage. Historical match data are stored as CSV files in S3 and versioned via Git. Data are loaded into SageMaker Studio for cleaning and feature generation.

Features: Rolling team strength (ELO/FIFA rank deltas), recent form (win/loss over k matches), head-to-head results, home advantage, rest days, and competition type.

We use Logistic Regression as the baseline, XGBoost for improved predictive power, and time-split train/validation/test for realistic forecasting.

1. **Data Sources**

Source: Kaggle International Football Results <https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017>

Volume: 40,000 matches, spanning more than 150 years

Rationale: Rich historical coverage, well-structured, suitable for classification tasks.

Risks: Class imbalance (fewer draws), era differences (rule changes, tournament structures), but no PII or PHI concerns.

Why selecting dataset: We chose this dataset because it provides a comprehensive history of international football matches, including qualifiers and tournaments. It is well-structured, reliable, and directly supports building models for outcome prediction. With the FIFA World Cup happening next year, this dataset is especially relevant for generating insights and forecasts ahead of the tournament.

1. **Data Engineering**

Storage: CSV files in S3, processed in SageMaker Studio.

Preprocessing: Cleaning missing values, standardizing team names, one-hot encoding categorical variables, date parsing, ELO feature computation.

1. **Training Data**

Split: Chronological (70% Train / 15% Validation / 15% Test) to mimic real forecasting.

Labeling: Pre-labeled (match outcome). No additional labeling required.

1. **Feature Engineering**

Fields Used: *home\_team, away\_team, home\_score, away\_score, date, venue, tournament.*

Transformations: One-hot encoding, rolling averages, ELO deltas, head-to-head streaks, recent performance.

Excluded Fields: Text notes, free-form commentary.

1. **Model Training & Evaluation**

**Algorithms:** Logistic Regression (baseline), XGBoost (final)

**Hyperparameters:** learning\_rate=0.1, max\_depth=6, n\_estimators=200

**Evaluation:** Accuracy 61.4%, Macro-F1 0.59. Best performance on Home Win predictions, weakest on Draw due to imbalance.

**Final Model Results (XGBoost):**

* Accuracy: **61.4%**
* Macro-F1 Score: **0.59**
* Model outperformed logistic regression by ~5 percentage points.

**Confusion Matrix Insights:**

* Strongest performance in predicting Home Wins.
* Lower recall on Draw outcomes due to class imbalance.

1. **Model Deployment**

Instance Type: ml.m5.large

Mode: Batch inference for match slates; real-time optional for interactive scenarios.

1. **Model Monitoring**

Model: Sataus, Features, Model Type, Classes, Test Accuracy

Infrastructure: CloudWatch (CPU, RAM, latency, error rates).

Data: Schema validation with Great Expectations, missing rate monitoring, categorical drift.

1. **Model CI/CD**

Checkpoints: Feature selection: Training, Deployment, Evaluation.

Tests: Unit tests, data leakage checks, reproducibility validation, performance threshold gating.

1. **Security, Privacy & Risks**

No PHI or PII processed. No credit card data.

S3 Buckets: AAI540FinalProject (read/write).

Bias: Historical biases (e.g., home advantage, rule changes).

Ethical Considerations: Evolution of competition structure may affect fairness of historical comparisons.

1. **Future Work**

* Incorporate **“ruleset era” features** (e.g., pre-WW2, post-WW2, modern era) to help the model account for shifts in match dynamics and improve prediction accuracy.
* Incorporate more granular features (e.g., player stats, weather, referee).
* Experiment with advanced models (e.g., neural networks, calibrated ensembles).
* Add automated retraining and model monitoring in production.
* Improve draw prediction through class rebalancing or SMOTE.
* Build a simple front-end for match forecasting.

1. **References**

Martj42. International Football Results from 1872 to 2017 (Kaggle Dataset). Retrieved from <https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017>

Amazon SageMaker. Machine Learning Service. Amazon Web Services.

GitHub. Version Control System.

XGBoost. Extreme Gradient Boosting.

scikit-learn. Machine Learning Library for Python.