

CSCI 538 – Artificial Intelligence Project Proposal

Title:

Predicting Athletic Department Financial Efficiency Using Two-Stage Machine Learning: A Hurdle Model Approach with XGBoost and SHAP Interpretability

Group Members:

Mukesh Ravichandran, Abner Lusung

Master of Science in Computer Science, East Texas A&M University (Fall 2025)

Problem Statement

Across U.S. universities, athletic departments manage more than \$18 billion annually, yet most lack analytical tools to assess whether their spending produces sustainable outcomes. Existing research focuses primarily on NCAA Division I programs, leaving mid-sized and smaller institutions without evidence-based guidance. **The federal Equity in Athletics Data Analysis (EADA) dataset** shows that a large share of institutions report balanced or surplus budgets (efficiency ≥ 1.0) in publicly disclosed data. According to the NCAA's own guidance, 'many, if not most' athletic departments report break-even results because institutional support absorbs deficits. This results in a **zero-inflated distribution** of efficiency ratios, complicating standard regression modeling. The central research question is: How can artificial intelligence predict athletic-department efficiency and uncover the structural and operational factors that most influence fiscal surplus?

Data and Data Collection

Data Source: U.S. Department of Education – Equity in Athletics Data Analysis (EADA) Portal (<https://ope.ed.gov/athletics/>).

Coverage: 2013–2023 (~17,200 institution-years from ~1,700 institutions).

All monetary values are inflation-adjusted to 2023 USD using the Consumer Price Index for All Urban Consumers (CPI-U) from the U.S. Bureau of Labor Statistics. The dataset contains institution-level data on revenues, expenses, athletic aid, salaries, recruiting costs, participation counts, and institutional context (Division, State, and Enrollment).

From **2,040 institutions** reported in the EADA dataset (2013–2023), only those with **complete ten-year data** were retained to support longitudinal modeling. Some institutions reported multiple times per year for different divisions, football status, or campuses—these were handled using the official **UNITID** identifier to ensure consistent tracking. The final dataset contains **1,722 unique UNITID-based institutions** with full reporting across all ten years.

Target Variable: Efficiency Ratio = Grand Total Revenue ÷ Grand Total Expenses

(range: [1.0, 4.9])

Predictor Variables (20 Features)

Category	Variable Name	Description	Type	Example
Financial	Grand_Total_Revenue	Total athletic revenue (2023 USD)	Float	12,340,000
Financial	Grand_Total_Expenses	Total athletic expenses (2023 USD)	Float	12,000,000
Financial	Total_Athletic_Aid	Total scholarships and grants (2023 USD)	Float	2,500,000
Financial	Total_Recruiting_Expenses	Recruiting budget (2023 USD)	Float	150,000
Financial	Total_Coaching_Salaries	Combined head + assistant coach pay (2023 USD)	Float	1,800,000
Financial	Men's_Team_Revenue	Men's team revenue	Float	5,600,000
Financial	Women's_Team_Revenue	Women's team revenue	Float	4,900,000
Financial	Not_Allocated_by_Sex_Sport_Revenue	Institutional support / student fees	Float	1,840,000
Participation	Total_Unduplicated_Athletes	Distinct athlete count	Integer	350
Participation	Male_Athletes	Male athlete count	Integer	180
Participation	Female_Athletes	Female athlete	Integer	170

		count		
Participation	Total_Undergraduates	Institutional enrollment	Integer	12,000
Participation	Women_Share	Female Athletes ÷ Total Athletes (%)	Float	0.486
Participation	Athletes_per_Undergrad	Athletes ÷ Undergraduates (%)	Float	0.029
Institutional	Division	NCAA Division (I, II, III, NAIA)	Categorical	D2
Institutional	State	U.S. state (2-letter code)	Categorical	TX
Institutional	Survey_Year	Reporting year (ordinal)	Integer	2023
Derived	Spend_per_Athlete	Expenses ÷ Athletes (\$)	Float	3,428
Derived	Revenue_per_Athlete	Revenue ÷ Athletes (\$)	Float	3,527
Derived	Aid_per_Athlete	Aid ÷ Athletes (\$)	Float	7,142

Data Processing Steps

1. Merge annual EADA CSVs (2013–2023) by UNITID and OPE ID.
2. Clean numeric fields and apply CPI-U inflation adjustment.
3. Impute missing values (median for continuous, mode for categorical by Division).
4. Drop features with >30% missingness and verify Efficiency Ratio consistency.
5. Encode Division (one-hot), State (target encoding), Year (ordinal).
6. Normalize continuous variables (z-score normalization).
7. Use temporal holdout: Train 2013–2020, test 2021–2023.

Goal / Scope

The objective is to construct a two-stage machine learning framework that accurately predicts athletic-department efficiency while remaining interpretable.

Stage 1 uses XGBoost for binary classification to predict whether an institution achieves surplus (Efficiency > 1.0).

Stage 2 applies XGBoost regression to predict surplus magnitude for those institutions.

Both models will be interpreted **using SHAP (Shapley Additive Explanations)** to quantify each feature's impact on predictions.

A **case study on Texas A&M University–Commerce (TAMUC)** will analyze how its NCAA Division I reclassification (2018–2020) may have influenced its efficiency trajectory between 2013–2023.

Expected Results

The Stage 1 classifier is expected to achieve target metrics in the range of ROC-AUC \approx 0.78–0.83 and accuracy \approx 75%, based on prior studies using XGBoost on similar institutional datasets (Chen & Guestrin, 2016; Fulks, 2015).

The Stage 2 regressor aims for $R^2 \approx$ 0.70–0.75 with RMSE \leq 0.12 efficiency units, in line with observed variance in financial efficiency modeling.

These are performance targets, not empirical results, they represent expected ranges drawn from prior literature and comparable applications. Preliminary hypotheses: structural factors (Division, Total Undergraduates, Total Expenses) determine whether an institution achieves surplus, while operational factors (Spend_per_Athlete, Women_Share, Revenue_per_Athlete) influence surplus magnitude.

The final analysis will test these expectations and replace benchmarks with actual performance metrics once modeling is complete.

References

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
- NCAA (2019). NCAA Financial Database Glossary. https://ncaaorg.s3.amazonaws.com/research/Finances/2019RES_NCAA_Financial_Database_Glossary_20191107.pdf
- U.S. Department of Education (2023). Equity in Athletics Data Analysis (EADA). <https://ope.ed.gov/athletics/>
- Fulks, D. L. (2015). Revenues and Expenses of Intercollegiate Athletics Programs Report. National Collegiate Athletic Association.