BSAN 6060 Project Proposal:

NCAA Predictive Cost-Benefit Analysis

Spring 2024 **Team:** Henry Lissner

Xiomara Vidal Marquez

Tina Brauneck

# Problem Statement

College and university athletics programs can be extremely costly and leaders need to determine how much to fund, with no guarantee that an increase in funding will yield better performance. With limited resources, what should the school bolster? Recruitment? Coaching? Student athlete incentives? Our project will use publicly available expense and win records to generate a predictive cost-benefit analysis that can assist leadership in their decisions.

# Data Sources

Our team will collect three types of historical data:

1. NCAA win/loss records
2. College/university athletic expenditures, which may include:
   1. Coach Salary Data
   2. Recruitment Expenses
   3. Program Operating Expenditures
   4. Student Financial Aid
3. College/University demographic data, such as:
   1. Location
   2. Enrollments
   3. NCAA classification

We have identified two sources where this information is available:

1. NCAA Statistics
   1. This source contains win-loss records and other performance indicators.
   2. <https://stats.ncaa.org/rankings>
   3. Sample:

A screenshot of a computer

Description automatically generated

1. Equity in Athletics Disclosure Act (EADA) Survey
   1. This survey includes demographic data for the nation's universities, their NCAA classification, and expense data related to their athletics programs.
   2. <https://surveys.ope.ed.gov/athletics/#/>

The data from these two sources appears to be already reasonably clean and well formatted, so we expect cleaning to be relatively low effort. One challenge for the analysis is drawing comparisons over a varied group of schools. We plan to mitigate this by narrowing our scope to schools in the NCAA Division I without football classification. Schools in this classification are comparators for LMU and the results will be most relevant to LMU leadership. Comparisons may still need careful handling, as sports offerings differ from school to school. This is essentially a "missing for cause" data issue.

One other potential challenge is fragmentation. NCAA statistics are organized by sport, so we may need to union several data sets to get a complete win-loss record set. We also need to carefully consider how the names of sports and schools align so we can join tables accurately. Additionally, should we opt to supplement the data with additional metrics from other sources, we may encounter further issues of fragmentation and naming.

# Architecture

Building a relational database to house this data has several benefits, among them are:

* Flexibility; ease of querying the data
* Opportunity for future analysis

The database will include, at minimum, the following entities:

* Colleges
* Win-Loss Metrics
* Expenses

# Analysis and Visualization

Our analysis, presented in Tableau, will likely include:

* Dual-axis line graphs showing spending over time compared to win-loss ratios
* Stacked bar chart showing athletic spending by college vs. overall operating expenditures
* Pair plots
* Regression analysis (independent variable = wins-loss ratio; dependent variables = expenditures by category, and possibly location)
* Geographic heat maps of win-loss ratios

Our analysis seeks to answer the questions:

* Does spending impact performance?
* What types of expenditures specifically drive performance?
* Is location a factor?

We are open to modifying our approach as we explore the data further.