

Respecting Order on the Seed List: Ordinal Models for NCAA Tournament Seeding in the NET-Quadrant Era



ABSTRACT

This study examines whether ordinal models outperform continuous regression when predicting NCAA tournament seed lines. To conduct the experiment, our team will compare an ordered logit model against regularized (Elastic Net) and tree-based regressions (Gradient Boosting) using NET Rank, schedule strength, opponent quality, win percentages, and quadrant summaries. Models will be tested with leave-one-season-out validation on selected teams only. The contribution is to determine whether honoring the ordered nature of seeding yields more accurate and more defensible seed forecasts than treating the seed line as just another numeric regression target.

BUSINESS PROBLEM FRAMING

Predicting NCAA tournament seed lines is an important business problem because seeding directly affects matchups, bracket difficulty, and how teams are viewed into the tournament. Because of this, more accurate seed predictions would be valuable to sports analysts, media outlets, NCAA teams and coaching staffs, as well as fans and bettors who rely on tournament projections.

For this project, seed lines are ranked outcomes from 1 to 16, meaning the order matters. A team moving from a 3-seed to a 4-seed is a much smaller change than moving from a 3-seed to a 12-seed. That makes seeding a structured decision rather than a simple continuous value. The project therefore focuses on whether models that recognize this ordered structure can better reflect how real seeding decisions are made.

The overall goal is to improve both the accuracy and interpretability of seed predictions using historical team résumé data. If successful, this approach could provide a more data-driven way to understand NCAA seeding decisions and could also apply to other real-world business problems where outcomes are naturally ranked rather than purely numerical.

RQ: When predicting seed line (1-16), do ordinal models outperform continuous regression-type models in out-of-sample accuracy?

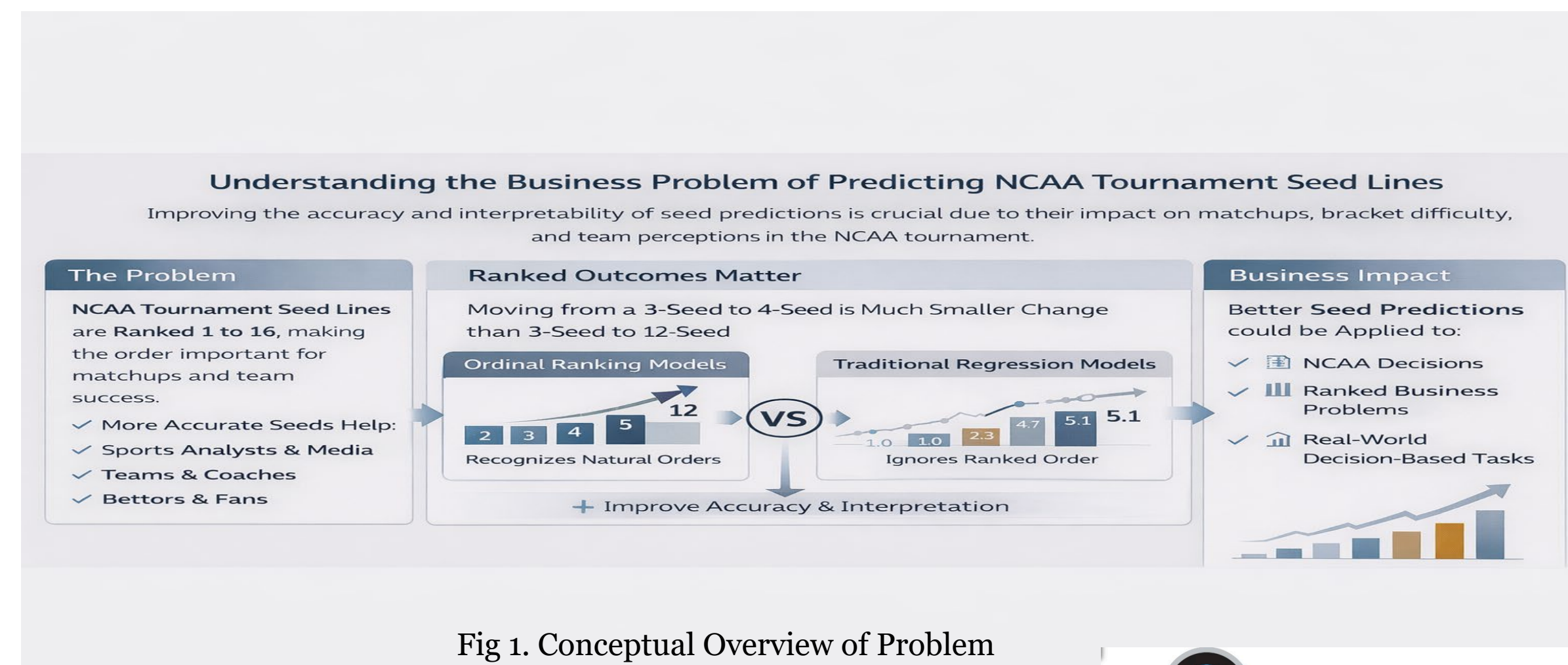
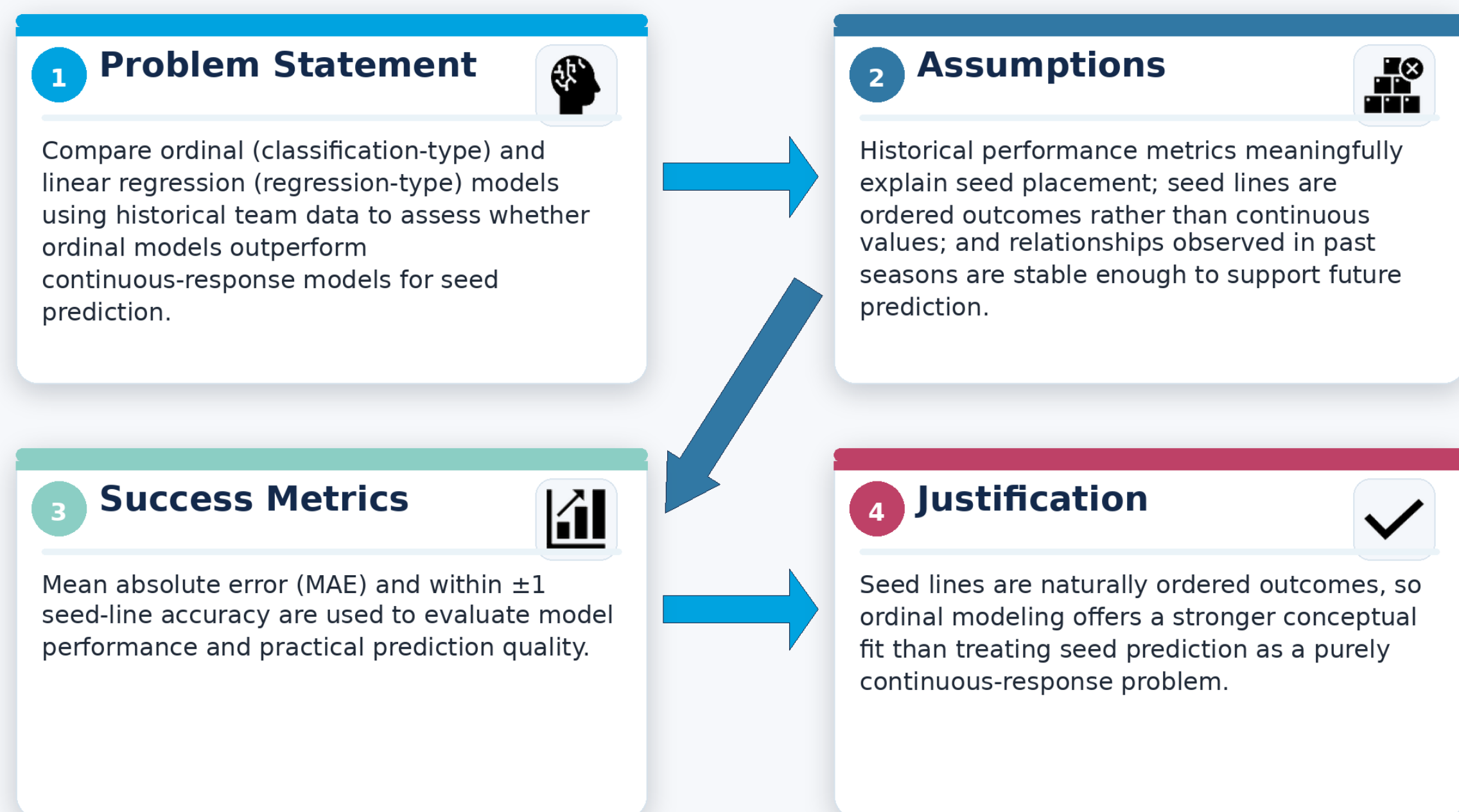
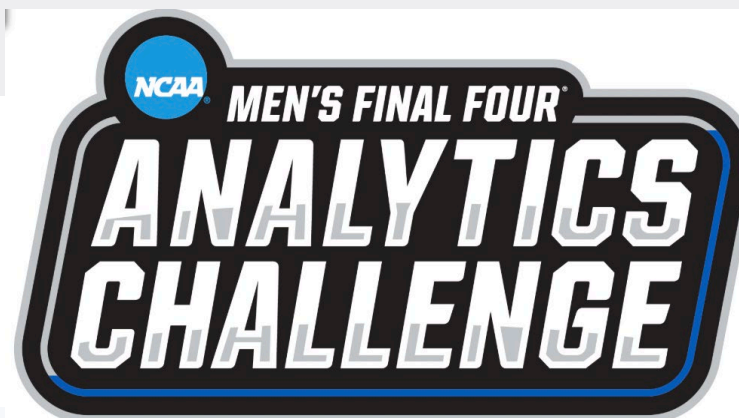


Fig 1. Conceptual Overview of Problem

ANALYTICS PROBLEM FRAMING



Personal Development & Outcomes

- Completed 5 DataCamp courses on AI Ethics, Power Pivot, Power Query, Understanding Excel, and Importing Data in Python
- Earned the Machine Learning Using SAS Viya Badge
- Developed functional Python literacy and data processing skills using Google Colab for the first time



DATA

Important Data Relationship – The NET Rank feature provides the biggest insights for seeding historically. In combination with Quadrant 1, and Wins/Losses, the data suggests that seeding follows a standardized metric-based approach.

Business Problem Framing Reflection – This process of data identification addresses the business problem by indicating key variables that when observed and tested provide the greatest insights into seed line ranking.

Feature Name	Data Type	Description	Sample Values
NET Rank	Integer	Measure of team quality using game results, opponent strength, and scoring margin.	1,9,47,112,289
PrevNET	Integer	Shows historical performance and year-over-year trajectory.	9,9,70,224,301
Overall Seed	Integer	Target variable – Seed number assigned by the NCAA (1= best, 16= weakest).	1,4,8,11,16
Quadrant 1 Wins/Losses	String	Record vs top-tier opponents. Q1 wins are the strongest signal while Q1 losses hurt seeding most.	"10-5", "0-0", "6-3", "2-4"

Table 1. Data Dictionary

METHODOLOGY

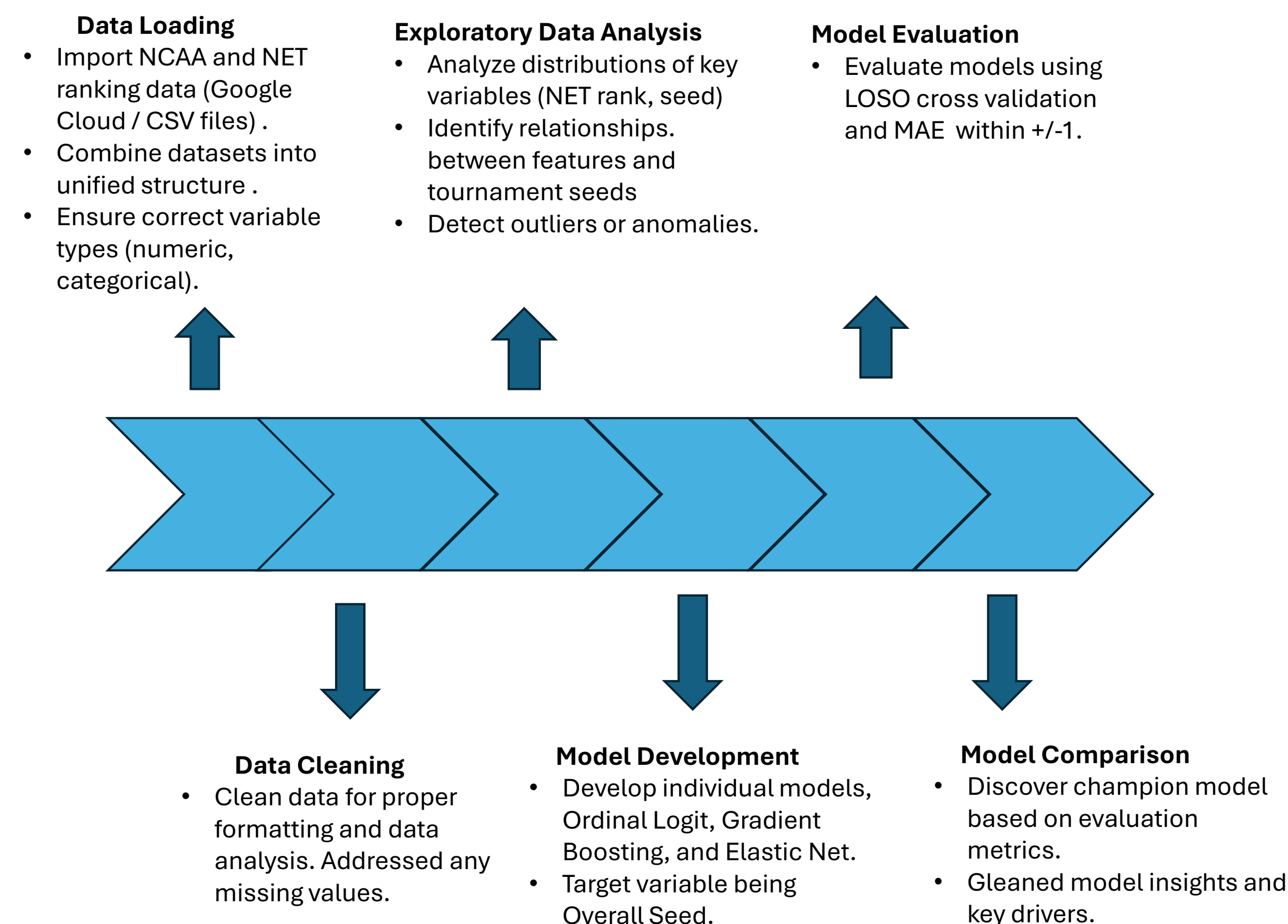


Fig 2. Methodology

This study follows a structured machine learning pipeline beginning with data collection and preprocessing, followed by exploratory analysis and feature selection. We used multiple models including Elastic Net, Gradient Boosting, and Ordinal Logistic Regression. Models were compared using Mean Absolute Error (MAE), with final selection based on test performance and generalization ability.

MODEL BUILDING & EXPERIMENTAL RESULTS

Table 2 summarizes performance of three model experiments. By comparing test and training statistics we can address models that overfit and generalize. All three are candidate (non-overfit) models.

Model	Train Accuracy	Test Accuracy	Train MAE	Test MAE
Elastic Net	99.96%	99.97%	1.21	1.15
Gradient Boosting	99.97%	99.95%	1.01	1.44
Ordered Logit	82.67%	82.34%	1.25	1.48

Table 2. Experiments

Model Performance Comparison (Training vs Validation MAE)

Lower MAE indicates better predictive accuracy.

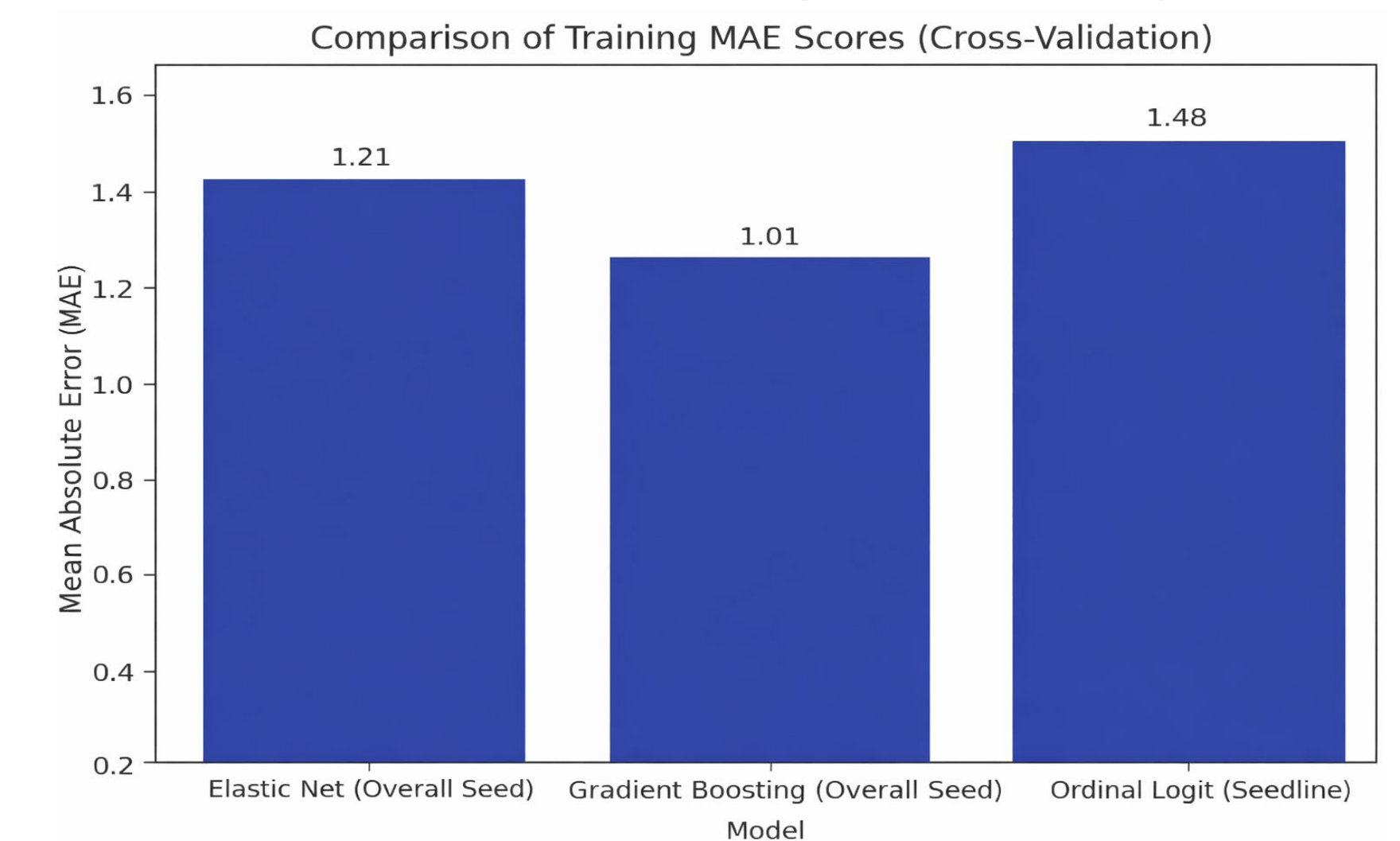


Fig 3. Experimental Results

Gradient Boosting achieved the lowest MAE, indicating the best predictive performance among our experiments

DEPLOYMENT & LIFECYCLE MANAGEMENT

- Among our experiments, gradient boosting was the champion model and suggests a machine-learning regression-based approach is preferred to ordinal ranking.
- Practically, our model experiments can be utilized to get a better understanding as to how the seeding process works, encouraging future exploration of ordinal models in the process.

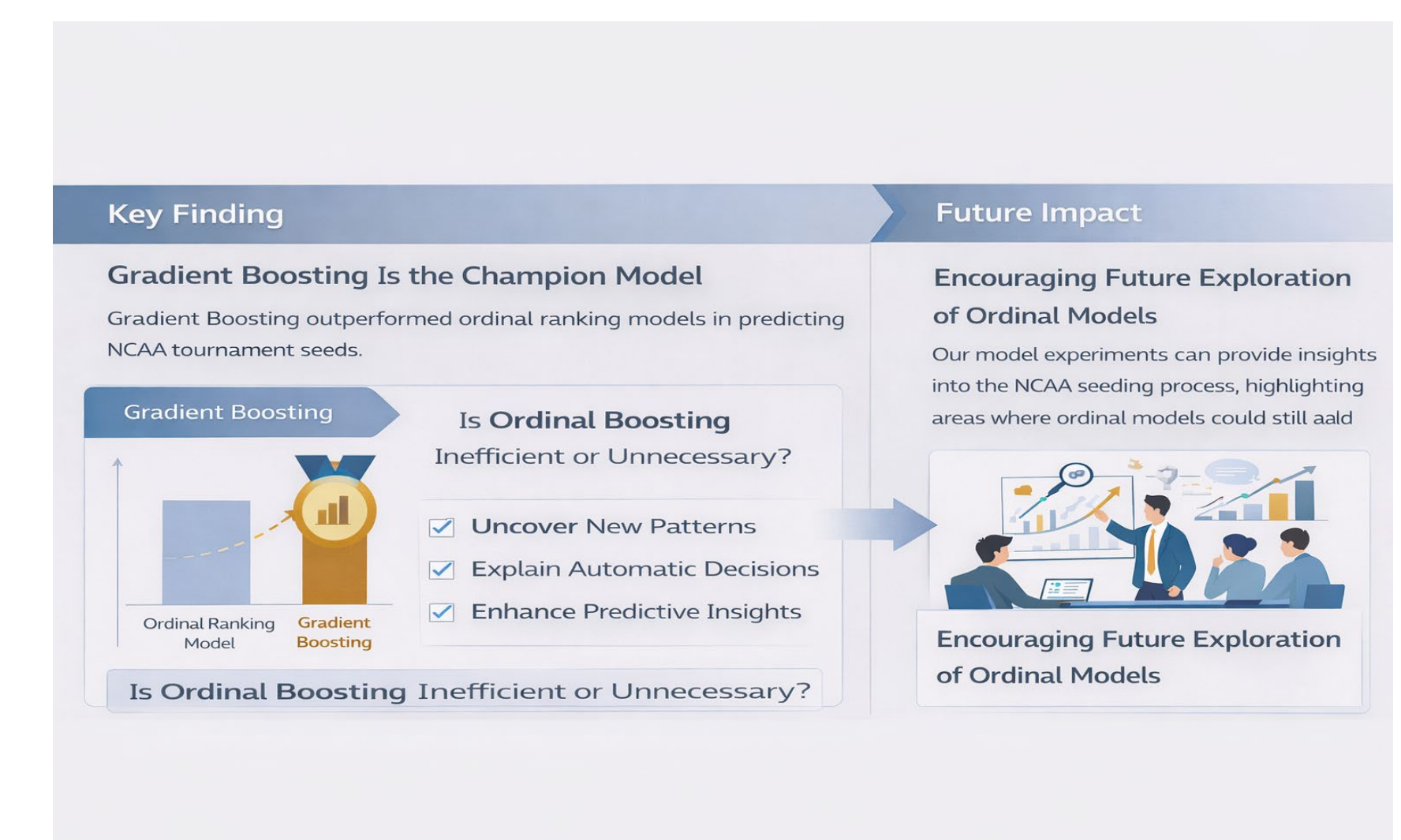


Fig 4. Deployment & Lifecycle Management Visualized

KEY TAKE-AWAYS

- We found continuous regression-type models outperformed ordinal models in out-of-sample accuracy, specifically regularized Elastic-Net models and Gradient Boosting decision-trees.
- We discovered the most important drivers for our models those being, NET Rank, Quadrant 1 Wins/Losses, and Prev Net.

Future Work

- While all our models were generalizable, we should probably spend time examining the recommended predicted seeds compared to the actual seeds to see if any unexpected results arise.
- We also would like to see if predictive performance is consistent across seeds (low vs. high).

Team Placement

- Our team placed 76th in the 2026 Final Four Analytics Challenge.

