

From At-Large Probability to Seed Line: A Two-Stage Framework for NCAA Tournament Modeling



Nik Belski, Aaron Bobinski, Abby Carter, Claire Venisnik, Max Winders, Dr. Matthew Lanham (Advisor)
Butler University Lacy School of Business
nbelski@butler.edu; abobinski@butler.edu; amcarter2@butler.edu; cvenisnik@butler.edu; mhwinders@butler.edu



ABSTRACT

This project asks whether NCAA tournament forecasting is accurate when selection and seeding are modeled as decisions rather than one joint problem. The literature offers foundations: Coleman et al. show a parsimonious at-large model can predict bids, whereas Reing and Horowitz demonstrate that selection and seeding can be formalized through a ranking framework. What remains underexplored is whether a staged pipeline mirrors how the committee works in practice. Using team-resume features such as NET Rank, schedule strength, opponent quality, win percentages, and quadrant results, we estimate a classification problem where response is selected/non-selected. Performance is evaluated with leave-one-season-out validation using AUC classification and adjusted R-squared regression metrics. Second, using the selections from stage one, each team's seeding is predicted. The expected result is a reproducible workflow connecting field construction to bracket placement and clarifying where predictive gains arise.

BUSINESS PROBLEM FRAMING

RQ: Does a two-stage modeling pipeline (selection first, then seeding) perform better than a single joint end-to-end model?

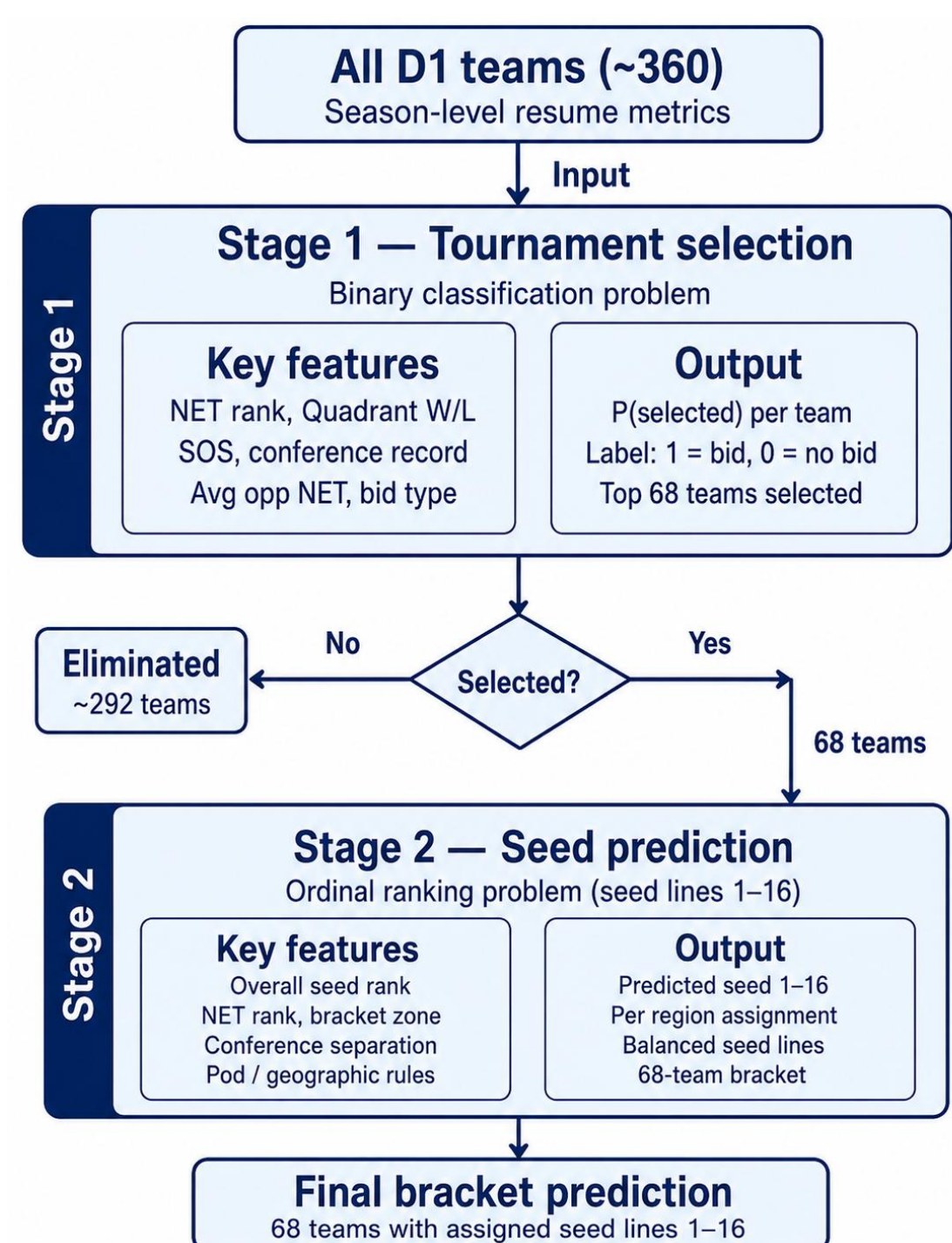
- The project predicts NCAA tournament bids and seedings using a two-stage approach:

- Stage 1:** Predict whether a team is selected.
- Stage 2:** Predict the seed of selected teams.

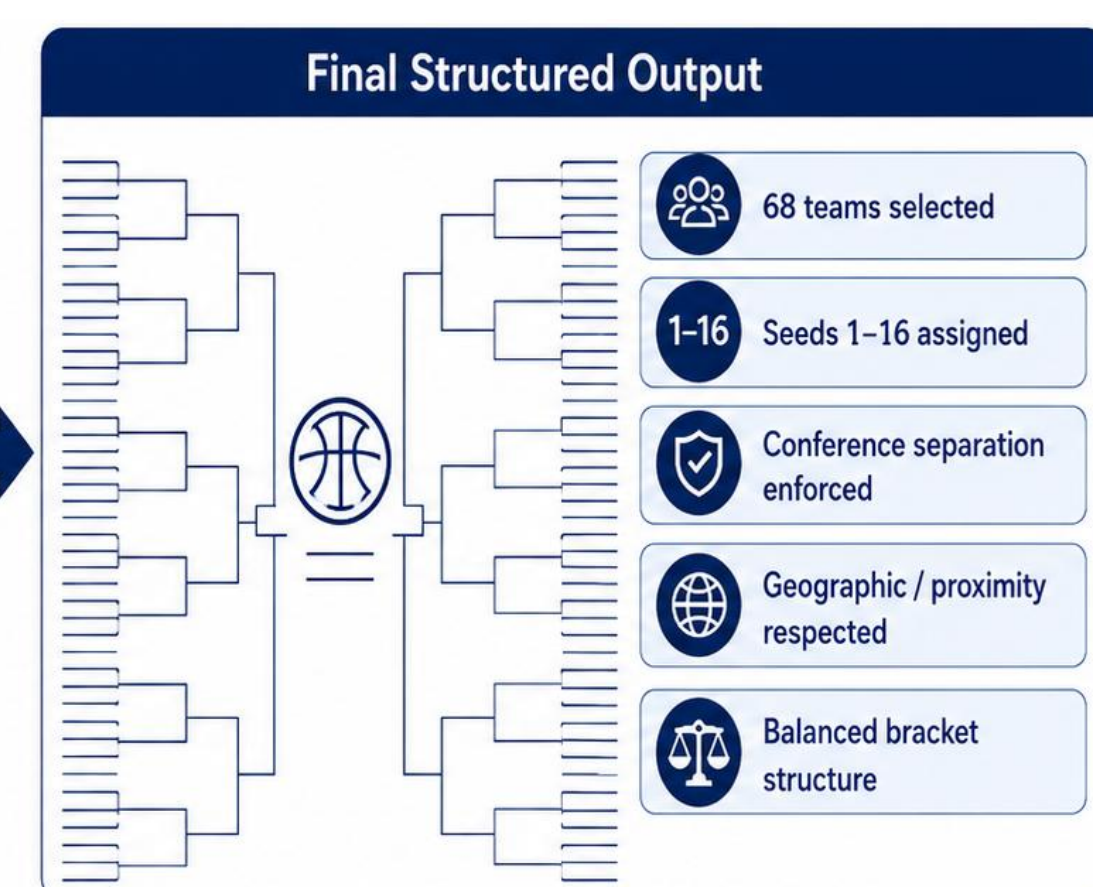
- This two-stage approach improves accuracy given the NCAA's complex and partially subjective selection process.

- The work benefits sports analytics researchers, data scientists, bracketologists, and fans interested in prediction.

- The model reflects real NCAA constraints—only 68 teams are selected, and seeding must follow rules on conference separation, geography, pod assignments, and balanced bracket structure.



Why a Two-Stage Approach?				
Model Type	Accuracy	Interpretability	Flexibility	Realism
Two-Stage Pipeline	High	High	High	Very High
Single End-to-End Model	Moderate-High	Moderate	Lower	Moderate



ANALYTICS PROBLEM FRAMING

- At-large selection is modeled as a binary classification problem using factors such as NET ranking, strength of schedule, Quad 1-2 wins, conference record, and road record.
- Assumes resume metrics align with NCAA committee criteria and that historical selection patterns are stable enough to model.
- Model performance is evaluated against historical results using metrics like AUC, with success defined by outperforming a baseline model.
- Prior research, including the Coleman study, supports using resume-based metrics for predicting tournament selection.
- The business problem centers on understanding and explaining NCAA selection decisions, while the analytics problem involves building a classification model to predict at-large bids from historical performance data.

Personal Development & Outcomes

- Completed 7 DataCamp courses on Beginner and Intermediate Python Programming, AI Ethics, and Understanding ChatGPT, Microsoft Copilot, Google Workspace with Gemini, and Excel.
- Coded Python in Google CoLab using Gemini assistant, leading to a greater understanding of how to prompt machine learning models.
- Earned Machine Learning Using SAS Viya Certificate.
- Presented at the Undergraduate Research Conference.
- Gained confidence from URC Conference prep and answering questions in class, even when not confident in my answers.

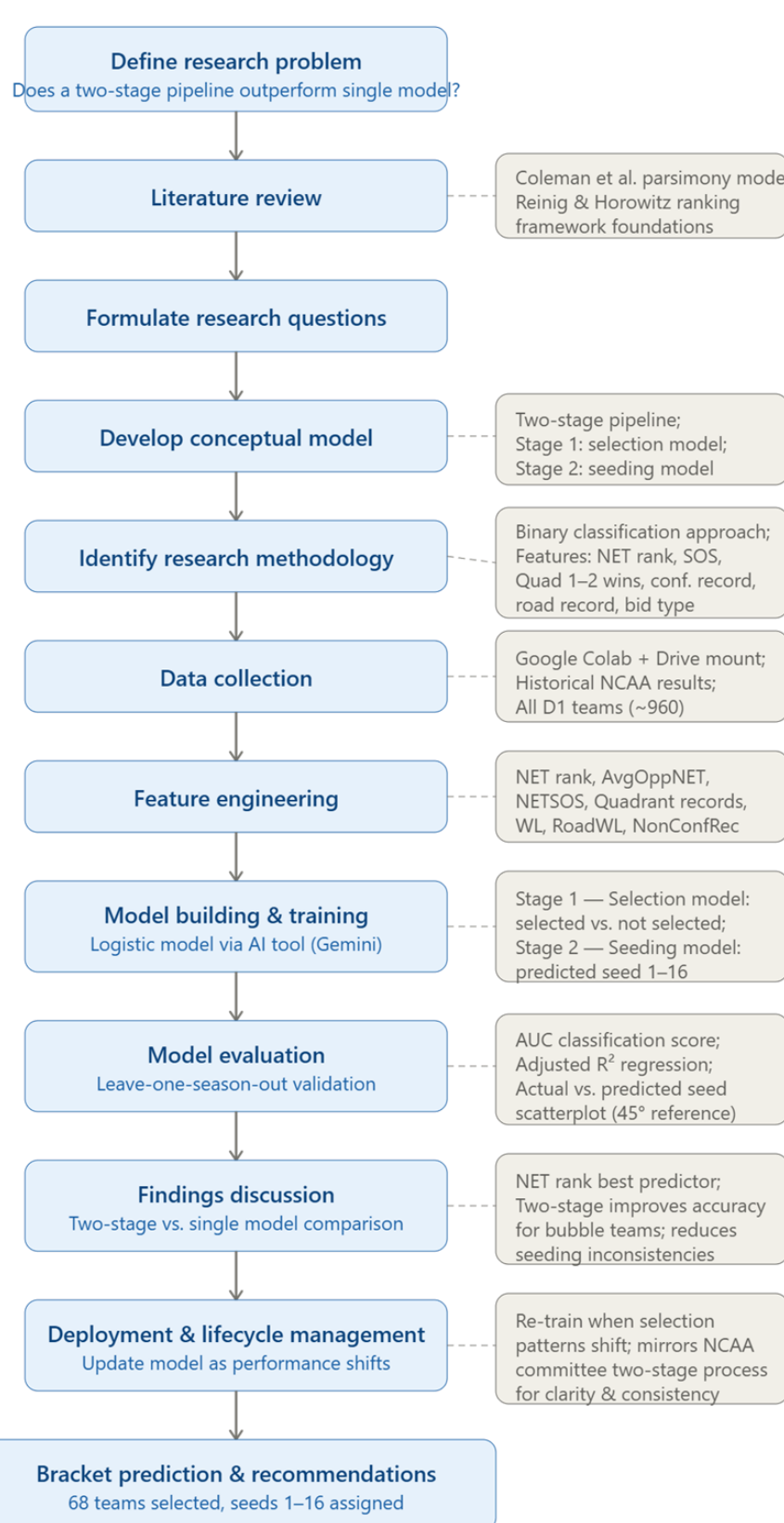
DATA

- Teams with stronger (lower) NET rankings are much more likely to receive at-large bids, reflecting the NCAA committee's reliance on NET as a key evaluation metric.

Team	NET Rank	Overall Seed	Bid Type
Baylor	2	1	AL
Alabama	2	9/24	AL/AQ
UConn	2	13	AQ
Arizona	2	7	AQ
Auburn	2	6	AL
Kansas	9	5	AL
Iowa St.	9	8	AL
Texas Tech	9	15	AL
Alabama	9	24	AL
Arkansas	14	15	AL
Kentucky	14	14	AL

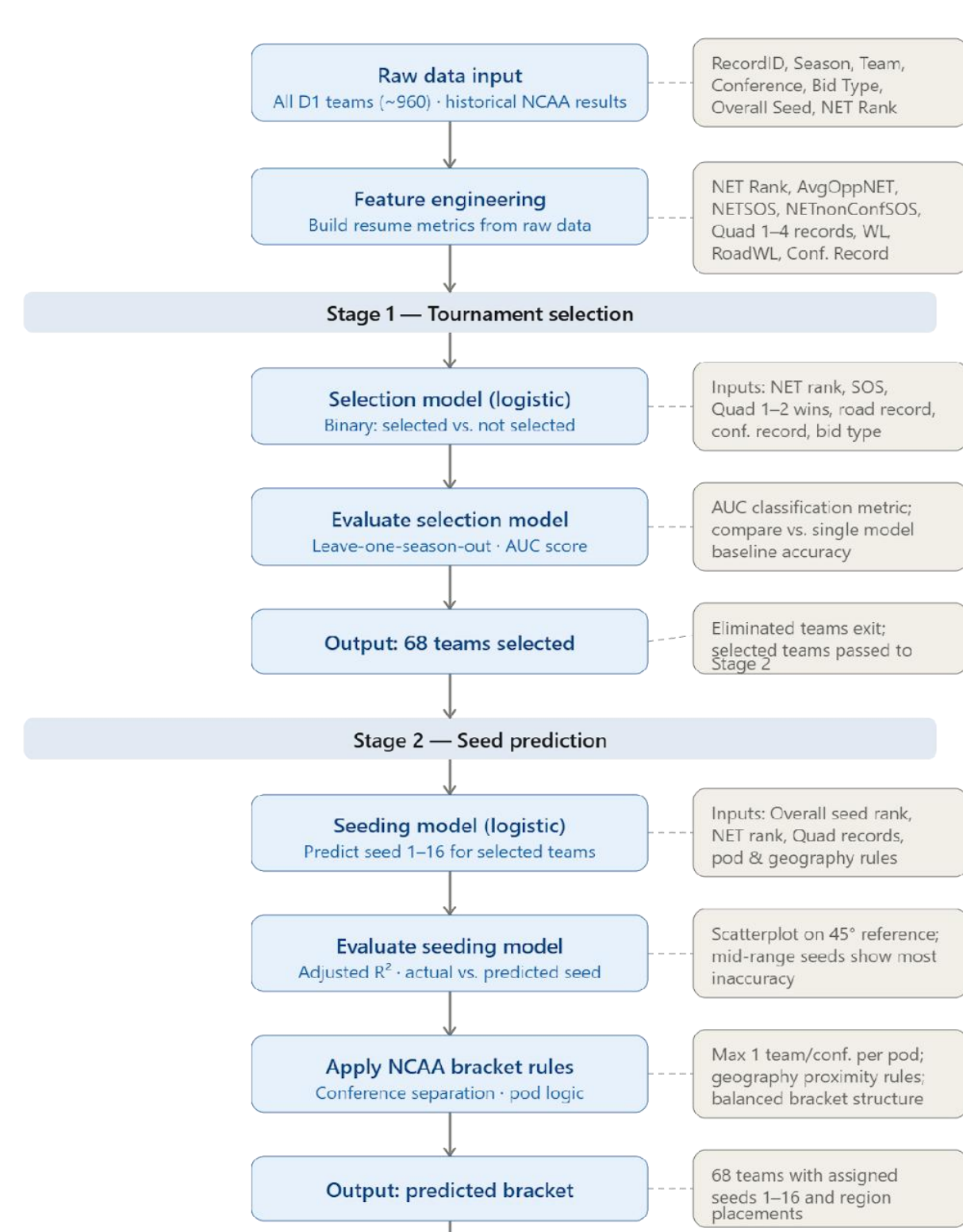
METHODOLOGY

Single Model Pipeline



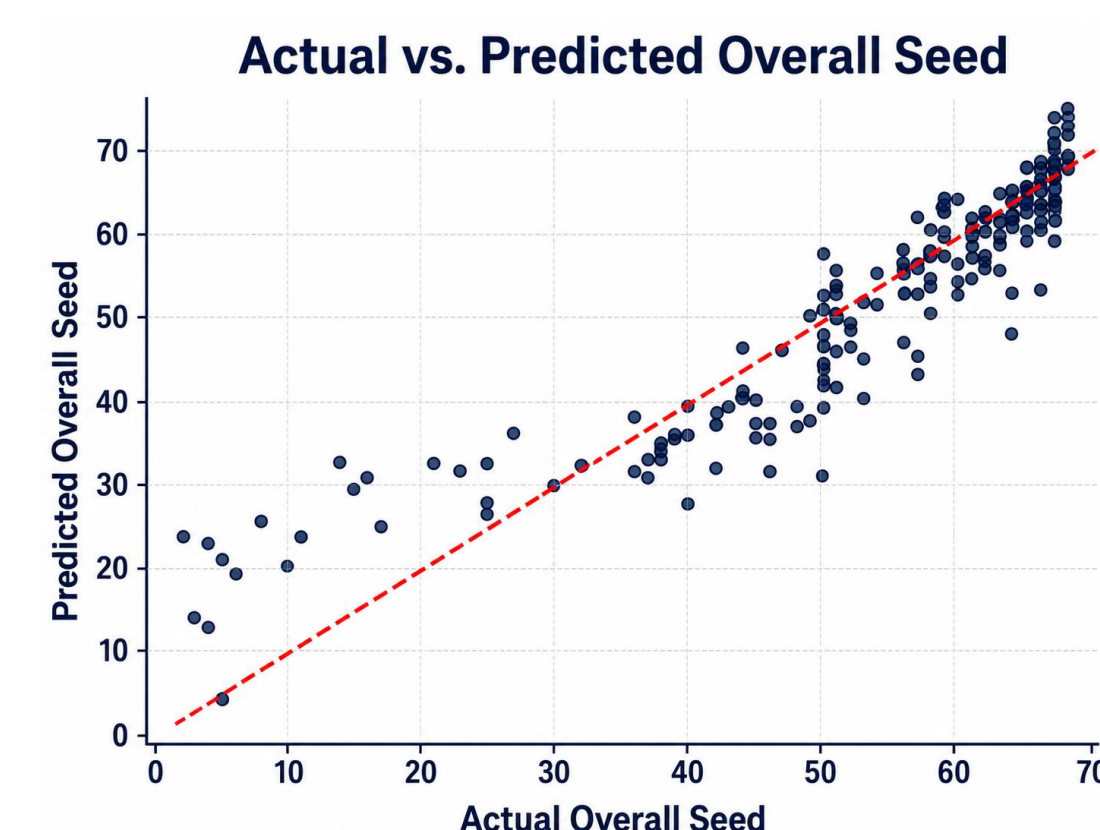
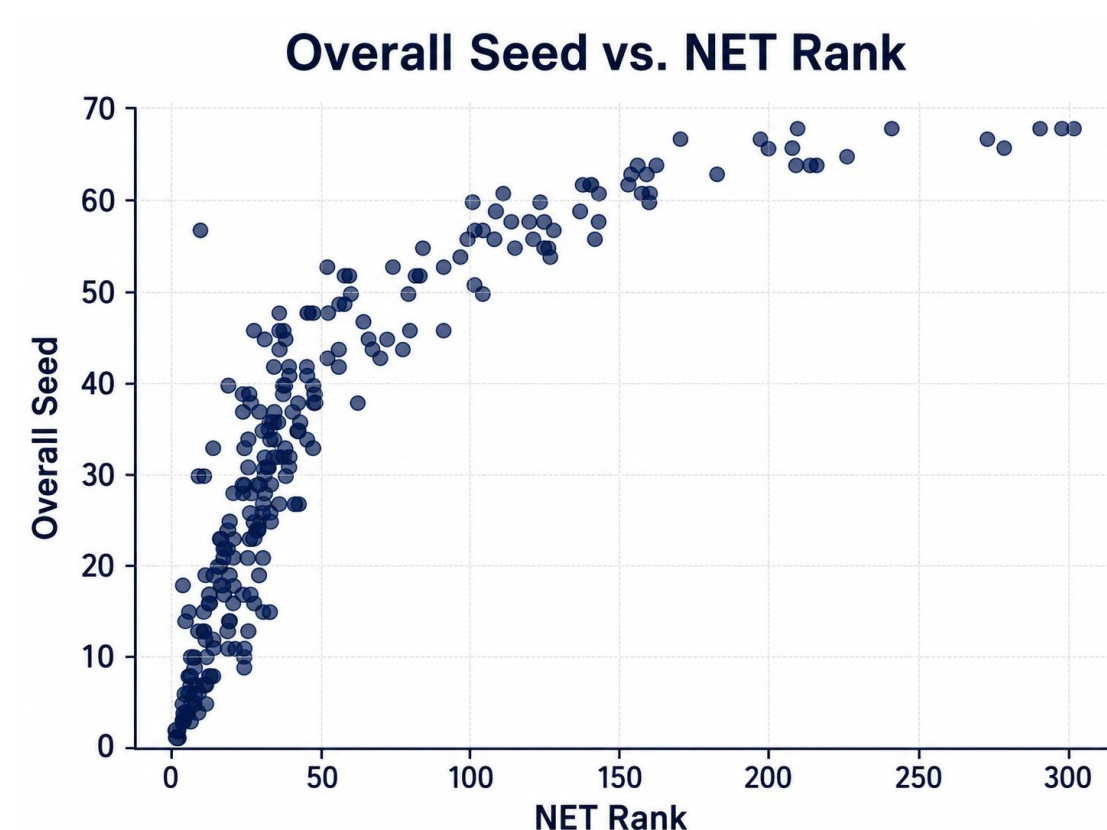
Feature	Description
RecordID	Unique identifier for each team-season record
Season	Academic year associated with the basketball season
Team	NCAA Division I school sponsoring men's basketball
Conference	Conference the team belongs to for that season
Overall Seed	Overall seed assigned to teams selected for the NCAA Tournament (1-68)
Bid Type	AQ = Automatic Qualifier (conference tournament winner), AL = At-Large bid selected by committee
NET Rank	NCAA Evaluation Tool ranking used by the selection committee to evaluate teams
PrevNET	Team's NET ranking from the previous ranking period
AvgOppNETRank	Ranking of the average opponent NET strength compared to other teams
AvgOppNET	Average NET ranking of all opponents played during the season
WL	Overall win-loss record
Conf.Record	Win-loss record against conference opponents
Non-ConferenceRecord	Win-loss record against non-conference opponents
RoadWL	Win-loss record in away games
NETSOS	NET Strength of Schedule ranking (1 = hardest schedule, 364 = easiest)
NETNonConfSOS	NET Strength of Schedule for non-conference games
Quadrant1	Record against highest-quality teams (top opponents based on location-adjusted NET rankings)
Quadrant2	Record against high-quality teams
Quadrant3	Record against mid-level teams
Quadrant4	Record against lowest-ranked teams

Two-Stage Pipeline



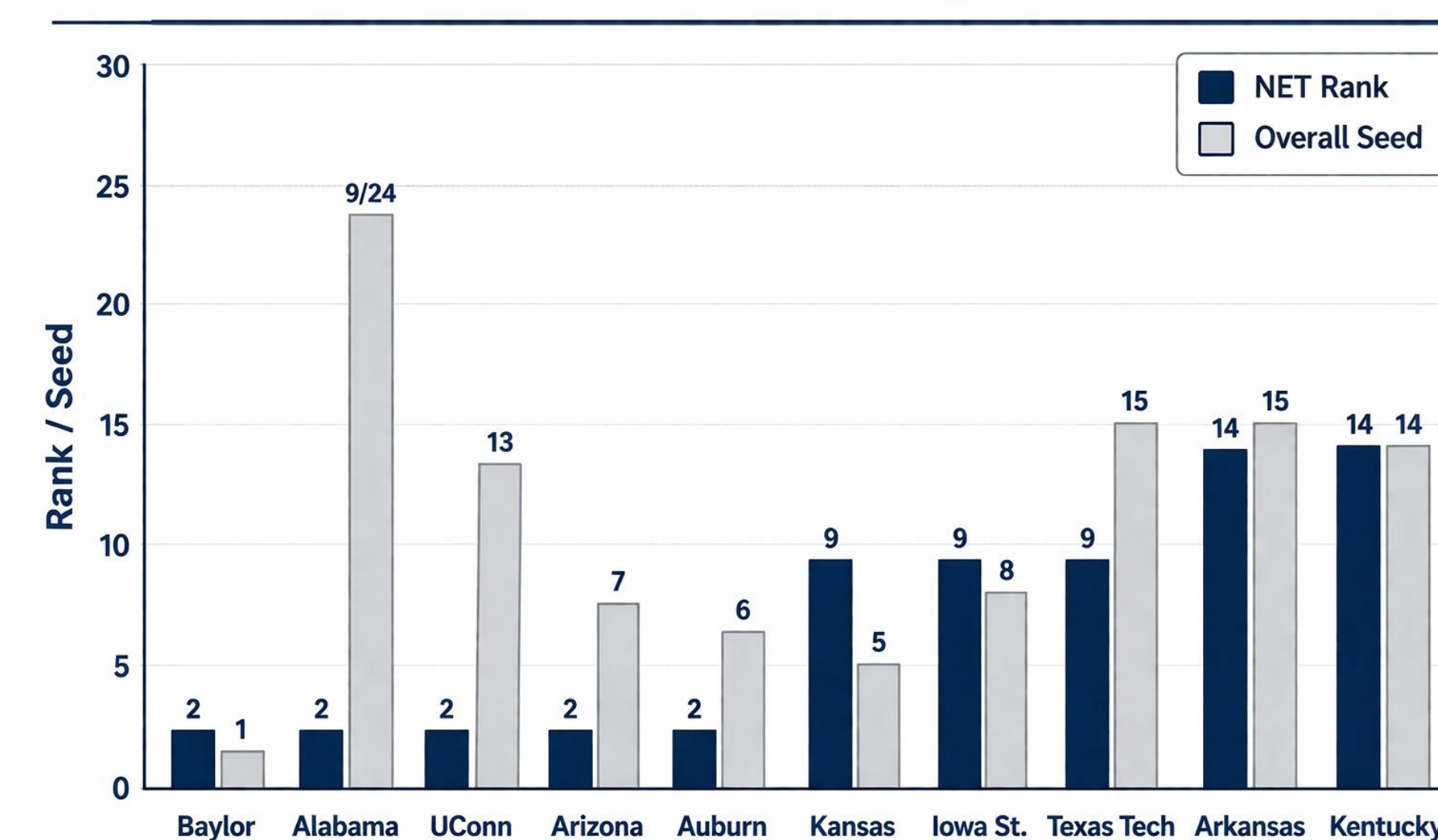
MODEL BUILDING & EXPERIMENTAL RESULTS

- We observed a strong relationship between actual and predicted seeds, with most points aligning closely to the 45-degree reference line; inaccuracies may occur mainly in the mid-range seeds.



- Besides a couple of outliers, the NET Rank was the best model for our data set. Both the test and training outcomes were similar for almost every team.

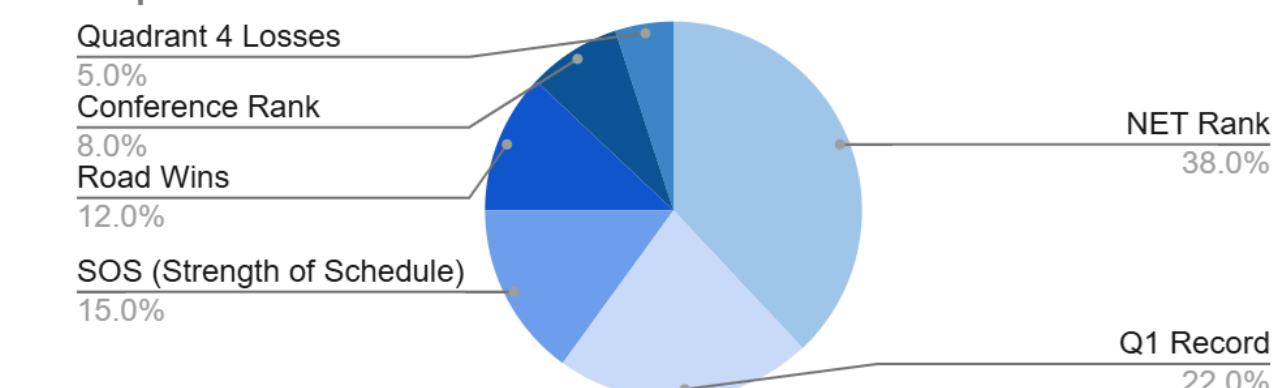
NET Rank vs. Overall Seed by NCAA Team



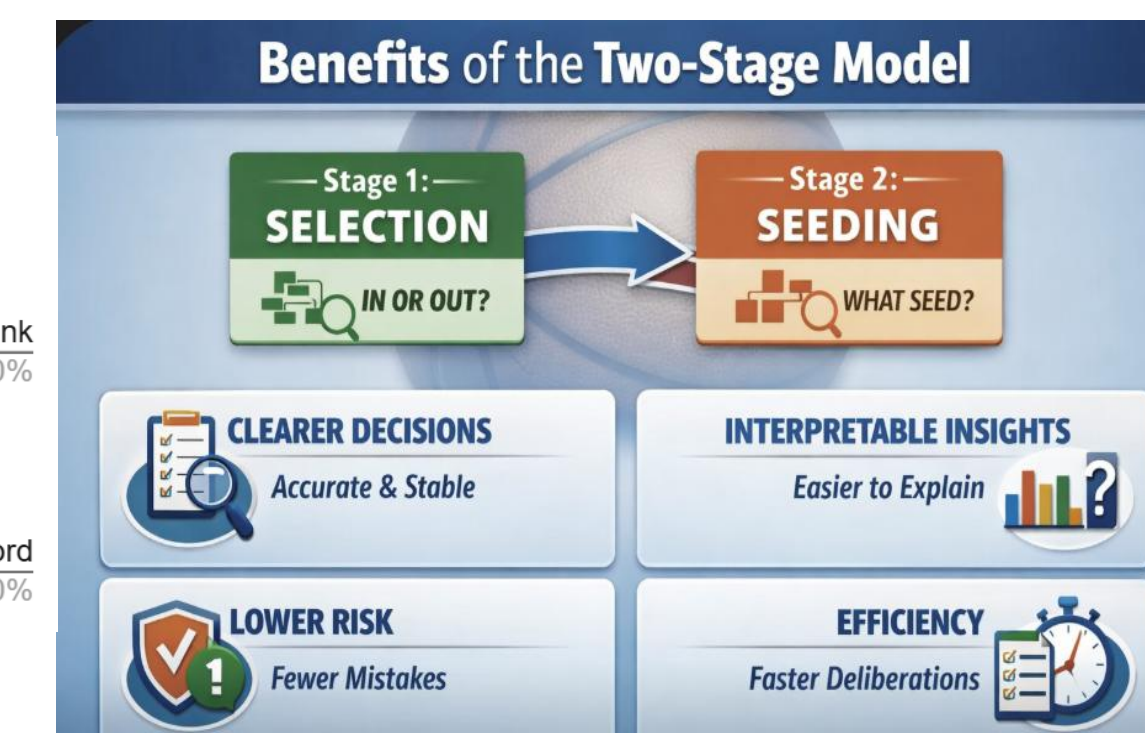
DEPLOYMENT & LIFECYCLE MANAGEMENT

- Separating selection from seeding improves accuracy, especially for bubble teams, and reduces inconsistencies that arise when a single model tries to predict everything at once.
- This approach also helps committee members justify decisions by clearly showing both selection likelihood and expected seed. The model can be updated as needed if performance declines or if relationships between variables change.
- Overall, our methodology mirrors the NCAA committee's process by using a two-stage model to enhance clarity and consistency.

Importance Score of Resume Metrics



The typical dominance of these resume metrics play a key role in future rankings done by NCAA committee members.



KEY TAKE-AWAYS

- The actual vs. predicted seed scatterplot showed points close to the regression line, indicating the two-stage model has solid predictive potential.
- The model did not fully predict all 68 seeds accurately, reflected in a lower-than-ideal R-squared value.
- Despite not scoring highly in the competition, the project provided meaningful experience with prediction methods, bracket-building factors, and hands-on model development.

