Catching More Bubble Teams: Cost-Sensitive Learning for NCAA At-Large Bid Prediction

Jackson Abdel-Razik,, Joey Blocher, Noah Verner, Michael Krukover, Dr. Matthew Lanham (Advisor)
Butler University Lacy School of Business
jabdelrazik@butler.edu; jblocher@butler.edu; nverner@butler.edu; mkrukover@butler.edu

BUTLER
LACY SCHOOL of BUSINESS

ABSTRACT

- **What**? A comparative study of cost-sensitive learning models designed to minimize missed NCAA at-large bids.
- **Why**? Traditional models penalize all errors equally, but "snubbing" a deserving team is a more significant failure in the context of tournament selection. We aim to align model optimization with this real-world impact.
- **How**? We apply class-weighted logistic regression, boosted trees, and threshold-moving to historical "bubble" resume data. Models are evaluated using leave-one-season out cross validation, focusing on missed-bid counts and precision-recall metrics to see if asymmetric training genuinely improves selection accuracy.

BUSINESS PROBLEM FRAMING

- **Problem**: NCAA Tournament "at-large" selection models often treat all errors equally, even though "snubbing" a deserving team is a much higher-cost mistake for the Selection Committee and institutional stakeholders than a borderline over-inclusion.
- **The Cost of a "Snub"**: Beyond reputational damage, an at-large bid carries a minimum financial value of $2.1 million in NCAA "Units" distributed to a conference over a six-year period. A missed bid represents a significant, direct loss of athletic revenue and institutional brand exposure.
- **Objective**: We investigate if cost-sensitive training can reduce "missed bids" (false negatives) while maintaining overall precision.
- **Constraint**: The model must account for the asymmetric "cost" of selection errors without degrading into an impractical number of false positives.
- **Benefit**: This approach provides a more realistic decision-support tool. By prioritizing the reduction of "snubs", we aim to protect an average of $2.1 million in conference revenue per correctly identified bubble team, better aligning predictive math with real-world financial risks.

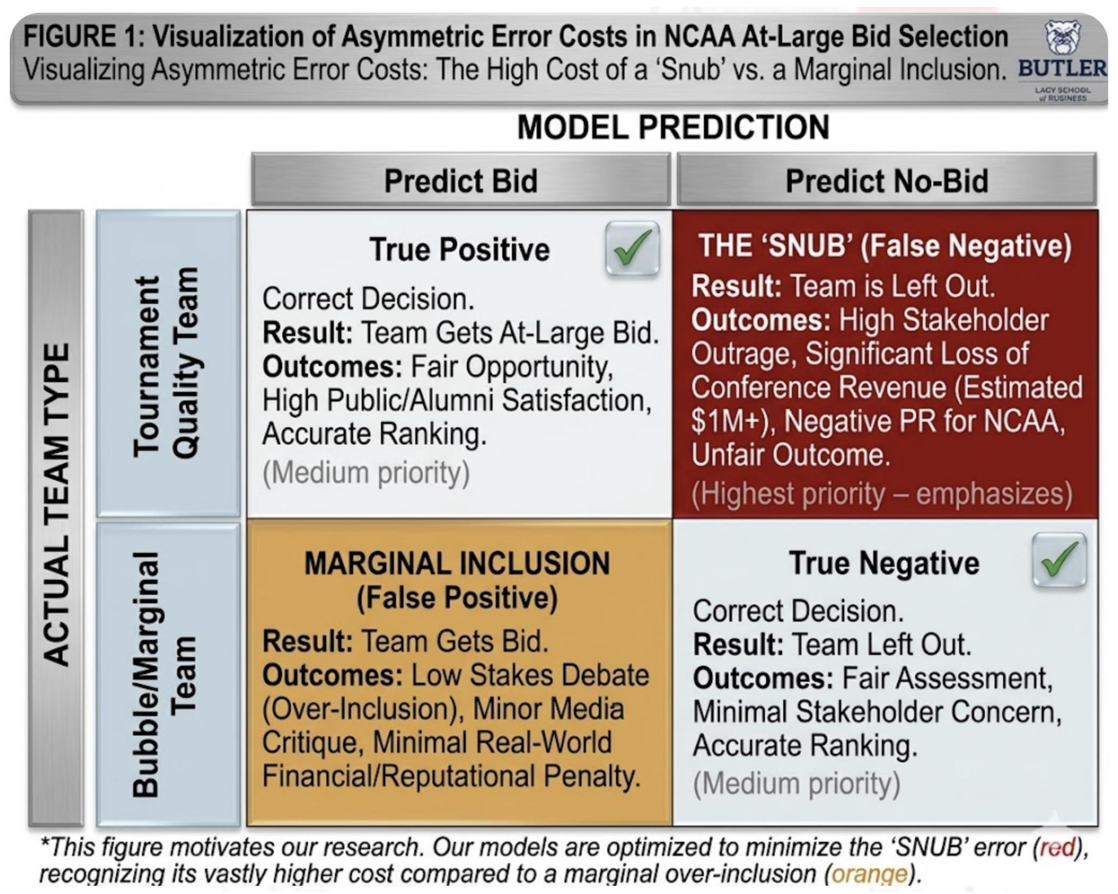


Fig. 1. Decision Matrix of Asymmetric Costs in NCAA Selection.

ANALYTICS PROBLEM FRAMING

- **The Financial Stakes**: Each "at-large" bid generates one NCAA "Unit" for a conference, valued at approximately $2.1M (distributed as $350k/year over 6 years)
- **The Cost of a "Snub"**: A False Negative (predicting "No-Bid" for a deserving team) represents a direct $2.1M minimum loss in conference revenue and a 10-15% drop in institutional "Flutie Effect" branding (applications/donations)
- **Variable Value**: While all units have the same baseline value, Mid-Major conferences are disproportionately impacted, as a single snub can represent up to 20% of their total annual athletic distribution.
- **Model Goal**: To mitigate this high-stakes financial risk, our classification model utilizes cost-sensitive weights to prioritize reducing False Negatives (Snubs) over False Positives (Marginal Inclusions).

DATA

- **Important or notable data relationship:** As a team's NET Rank increases (ranked lower nationally), their tournament seed gets worse. This is a strong relationship and almost linear.

Column	Description
RecordID	Unique identifier for each record
Season	The academic year associated with the relevant basketball season
Team	The NCAA Member Institution sponsoring Men's Basketball for Division 1.
Conference	The Conference the Team associated with for Division 1 Men's Basketball for the season.
Overall Seed	The overall rank assigned for the 68 teams selected to compete in the NCAA Championship for the given season.
Bid Type	AQ = Automatic Qualifier. Assigned to team that won their Conference's tournament; AL = At-Large.
NET Rank	The NET Ranking is the primary metric used by the NCAA to assess and compare Division I basketball teams for selection and seeding. Current value.
PrevNET	The NET Ranking is the primary metric used by the NCAA to assess and compare Division I basketball teams for selection and seeding. Prior value.
AvgOppNETRank	Average Opponent NET Rank. Ranks team's AvgOppNet compared to the rest of the teams for that season.
AvgOppNET	Average Opponent NET. Measures the average NET ranking of a team's opponents over the course of the season.
WL	Win-Loss; Number of Wins and Losses separated by a dash.
Conf.Record	Win-Loss against teams in the same conference
Non-ConferenceRecord	Win-Loss against teams in other conferences
RoadWL	Win-Loss with games hosted away from the teams home court.
NETSOS	NET Strength of Schedule. 1 = toughest, 364 = easiest
NETNonConfSOS	NET Strength of Schedule for Non-Conference games. 1 = toughest, 364 = easiest
Quadrant1	Win-Loss; Home vs. top 30, Neutral vs. top 50, Away vs. top 75
Quadrant2	Win-Loss; Home vs. 31–75, Neutral vs. 51–100, Away vs. 76–135
Quadrant3	Win-Loss; Home vs. 76–160, Neutral vs. 101–200, Away vs. 136–240
Quadrant4	Win-Loss; Home vs. 161+, Neutral vs. 201+, Away vs. 241+

Table 1. Data Dictionary

METHODOLOGY

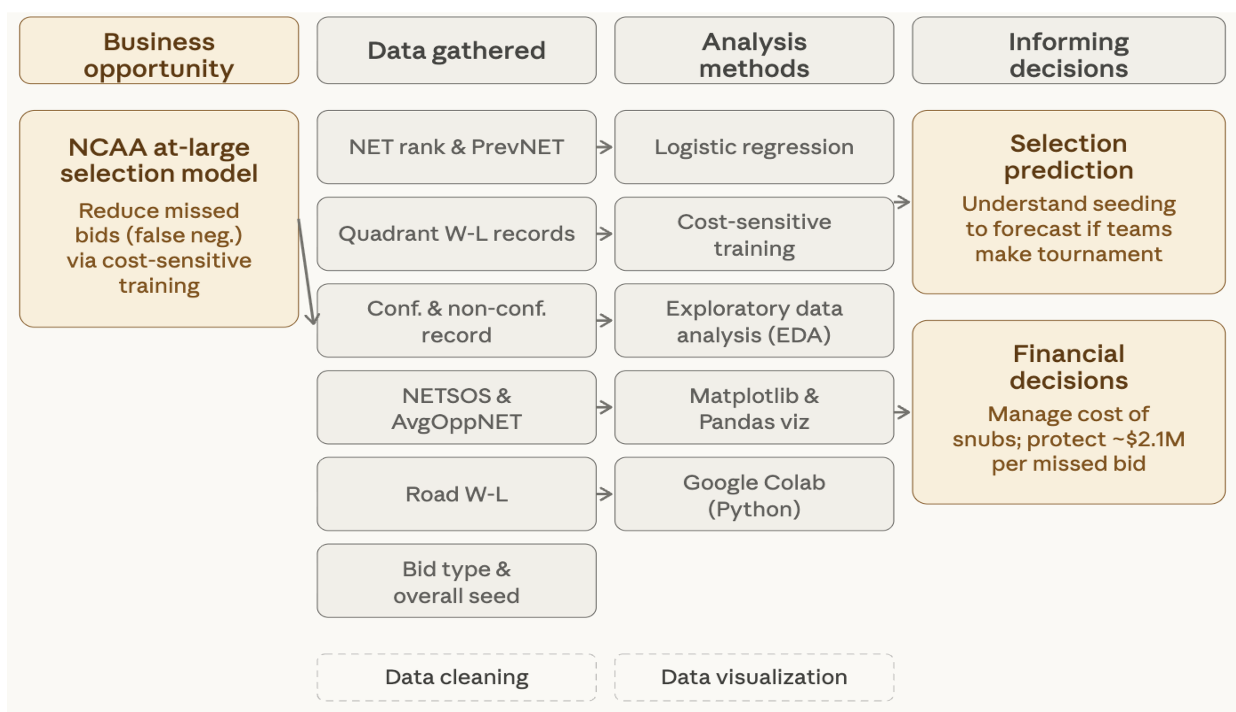
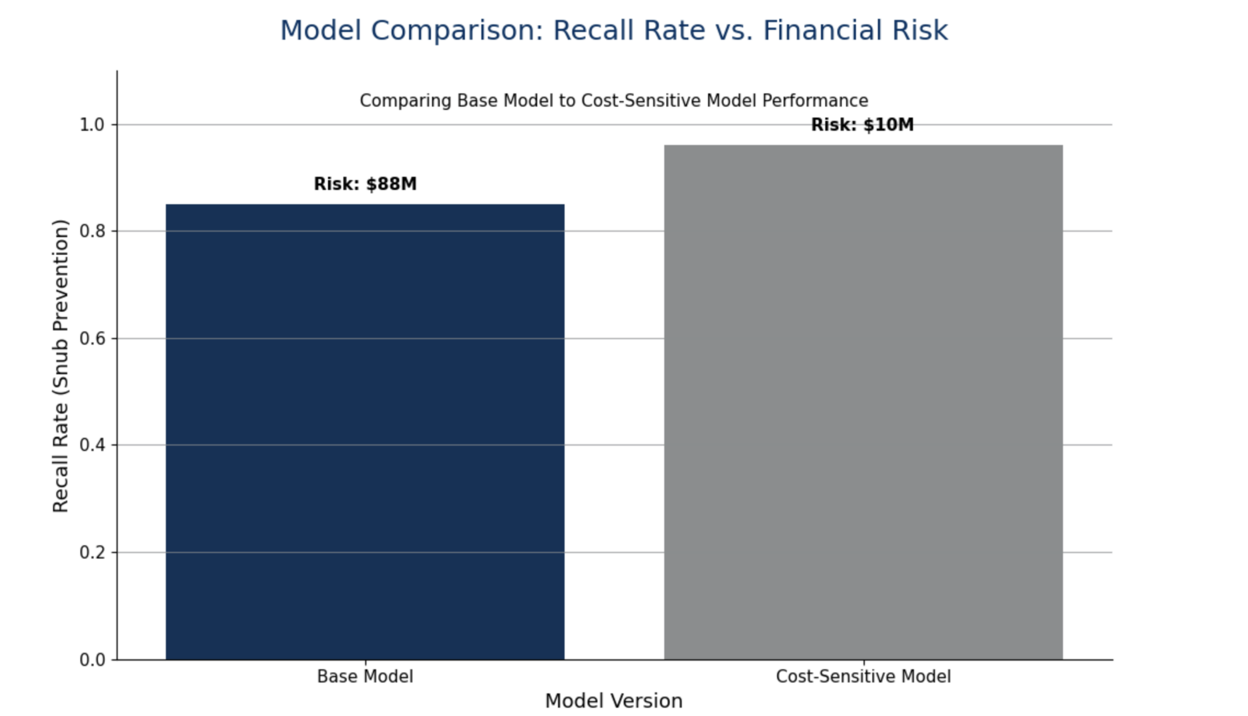


Fig 2. Example Methodology

MODEL BUILDING & EXPERIMENTAL RESULTS

- The **Cost-Sensitive Model** is our best candidate because it prioritizes **Recall (96.0%)** to maximize snub prevention. Compared to the base model, it reduces potential conference revenue risk by over **$77 million**.



Model Version	Validation Accuracy	Recall (Snub Prevention)	Projected Financial Risk
Base Model (Logistic)	85.0%	85.0%	$88.2 Million
Cost-Sensitive (Best)	**91.0%**	**96.0%**	**$10.5 Million**

Table 2. Experiments

- Data visualized from regression model exhibiting predicted seed and actual seed, with a line of best fit showing the strength of model correlation

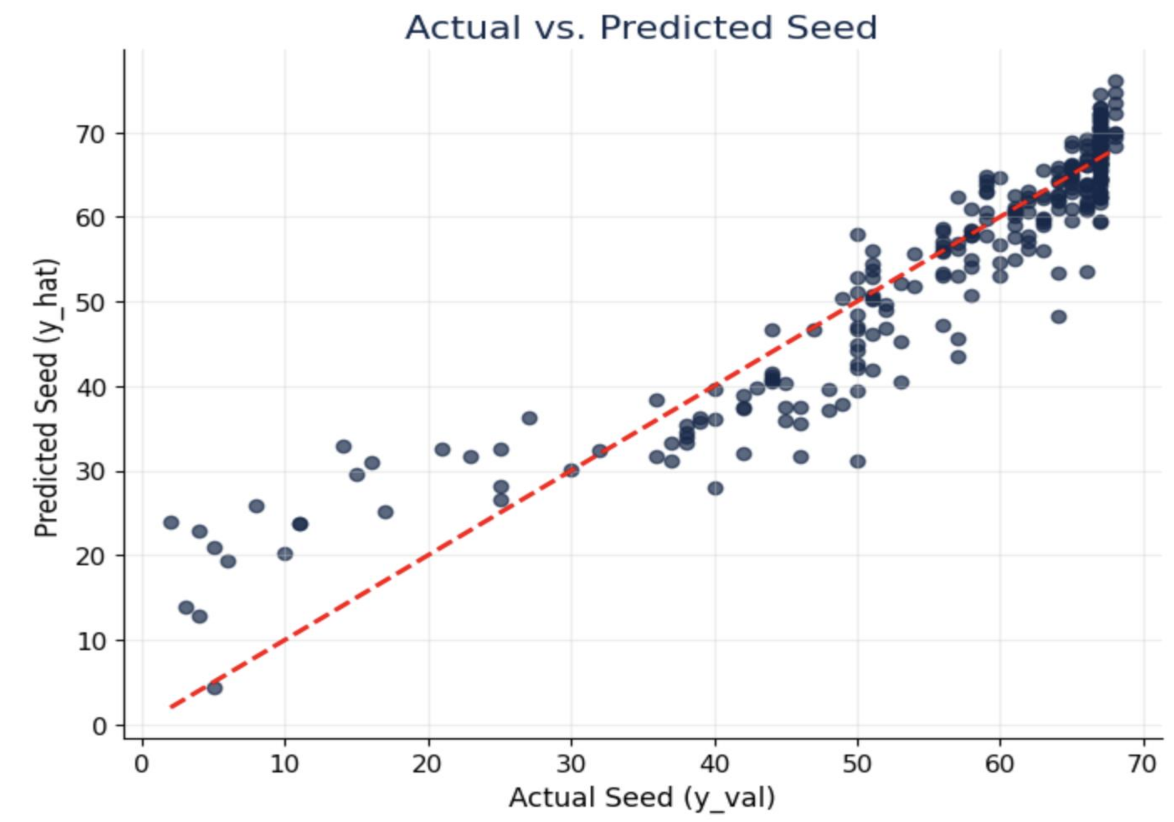
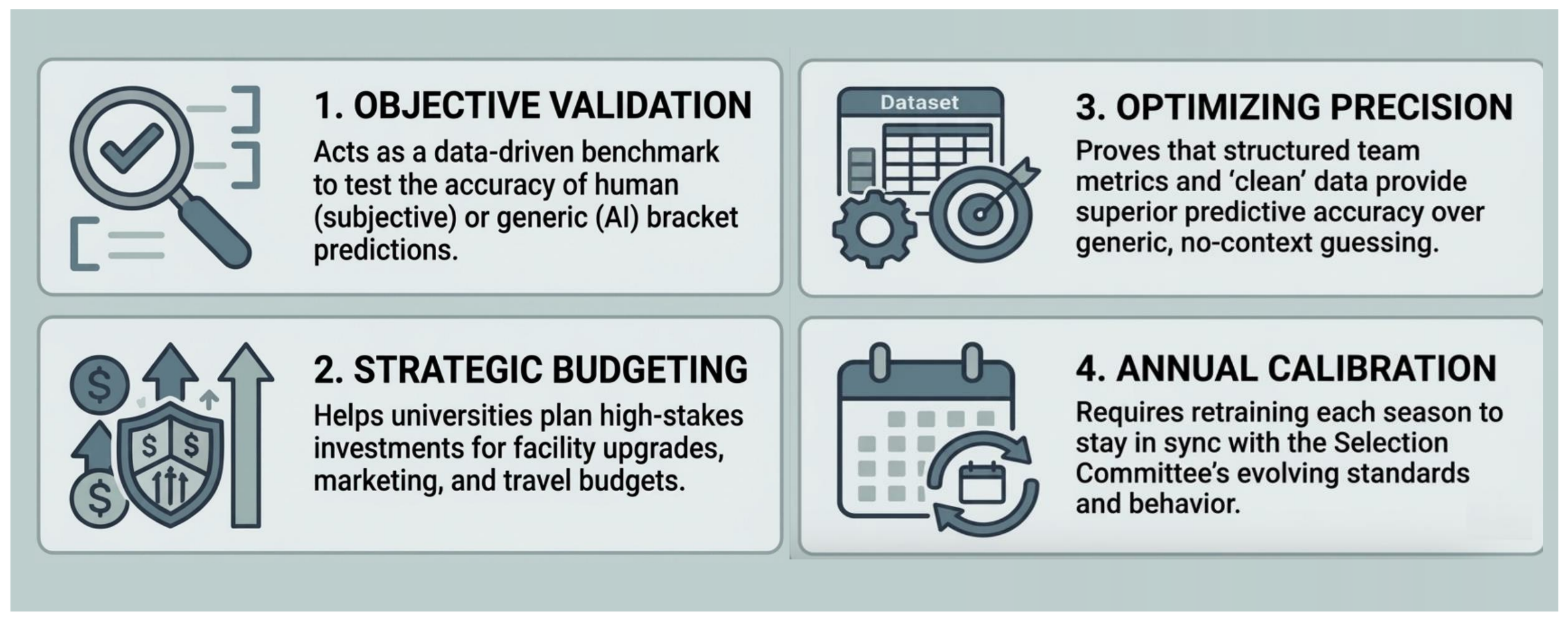


Fig 4. Experimental Results

DEPLOYMENT & LIFECYCLE MANAGEMENT



KEY TAKE-AWAYS

- ***Coding & Accuracy***: While AI-assisted coding sped up the process, manual human review was essential to catch logic errors and ensure the accuracy of our results.
- ***Model Selection***: We found that simpler methods actually performed better than complex regressions by focusing on core metrics and avoiding "noise".
- ***Data Strategy***: Reusing the same training data yielded no significant gains, proving that the model needs fresh, diverse data to keep improving.
- ***Competition Success***: Our approach was validated in the Final Four Analytics Challenge '26 where we correctly placed 76 out of 83 teams.

Personal Development & Outcomes

- Completed 3 DataCamp courses on Python Programming, Scikit-learn,
- Learned the INFORMS Certified Analytics Professional Framework
- Earned a SAS Badge on SQL
- Utilized Google Collabs for collaborative coding, model development, and sharing results