

How Sure Is the Bubble? Calibrated Selection Probabilities and Uncertainty-Aware Seed Forecasts



C. Gallagher, M. Malecki, J. Miller, B. Schnackenberg, M. Smith, Dr. Matthew Lanham (Advisor)
Butler University Lacy School of Business
cgallagher@butler.edu; mmalecki@butler.edu; jmiller@butler.edu; bschnack@butler.edu; msmith@butler.edu



ABSTRACT

This project examines whether NCAA tournament selection models can produce well-calibrated probabilities and reliable seed forecasts. Using KenPom data from 2020–2024 (N=500), we train logistic regression and gradient boosting models, then apply post-hoc calibration (Platt scaling, isotonic regression). Seed uncertainty is captured through conformal prediction intervals. Under leave-one-season-out validation, calibrated models improve probability trustworthiness, especially for bubble teams.

BUSINESS PROBLEM FRAMING

RQ: Does post-hoc calibration improve NCAA tournament selection probabilities, and can seed forecasts include meaningful uncertainty ranges?

Each March, the NCAA Selection Committee picks 68 teams. About 36 at-large bids are contested, creating a “bubble” where small differences determine who gets in. Analysts rely on models, but most give only yes/no predictions without expressing how confident those predictions really are.

A team with a 51% chance of selection is very different from one at 95%, yet traditional accuracy metrics treat both the same. Better-calibrated probabilities help analysts, coaches, and media understand the true level of uncertainty for each team.

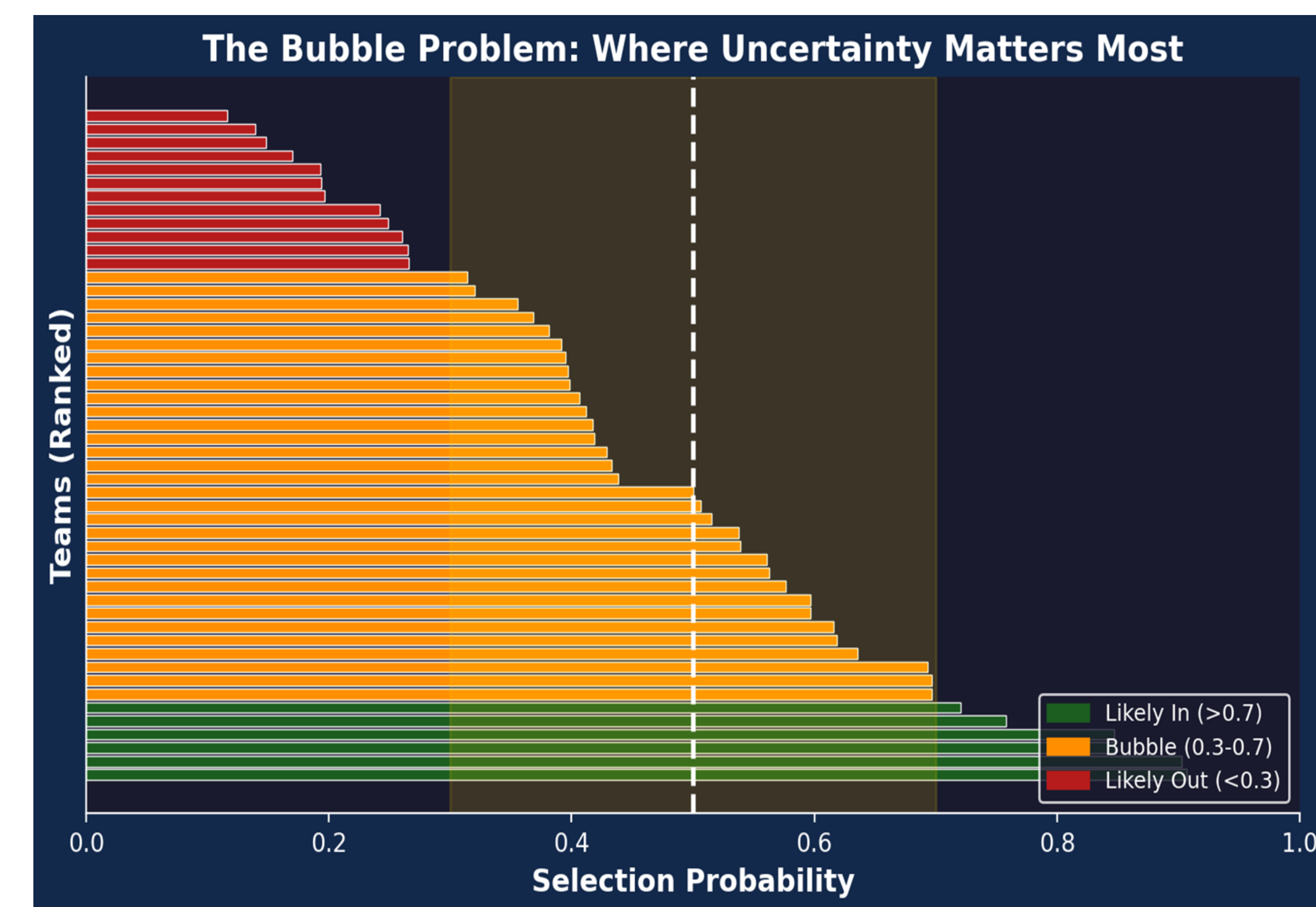


Fig 1. Conceptual Idea

ANALYTICS PROBLEM FRAMING

We address two related analytics problems:

(1) Classification: Predict whether a team makes the tournament (in vs. out) and produce calibrated probability estimates using post-hoc methods.

(1) Regression: Predict tournament seed for selected teams and construct prediction intervals that capture seed uncertainty.

Success is measured by Brier score, expected calibration error (ECE), reliability diagrams, and interval coverage—not just classification accuracy.

Personal Development & Outcomes

- Learned to develop statistical analysis using SAS systems, as well as ways to present said data.
- Learned how to work with a team on a analysis project.
- Used AI to develop code, which also developed my own skills on software.
- Learned how to present data and analysis in a way that can be presented to people, in a way that makes sense to them.

DATA

- **Source:** 5 seasons of data (2020–2024), covering ~100 teams per season (500 teams total). We used historical tournament brackets to label which teams actually made it.
- **Team Stats Used (7 total):** Net Rating, Offensive Rating, Defensive Rating, Win %, Strength of Schedule, Luck Score, and National Rank.
- **What we predicted:** (1) Did the team make the tournament? (yes/no) and (2) What seed did they get? (1–16). About 56% of teams in our dataset made the tournament. 281 teams that made it were used for seed prediction.

| Feature | Description | Type | Range |
|------------|-------------------------------|------------|---------------|
| NetRtg | Adjusted Efficiency Margin | Continuous | -5 to 37 |
| ORtg | Adjusted Offensive Efficiency | Continuous | 95 to 128 |
| DRtg | Adjusted Defensive Efficiency | Continuous | 85 to 108 |
| WinPct | Win Percentage | Continuous | 0.45 to 0.97 |
| SOS_NetRtg | Strength of Schedule (Net) | Continuous | -5 to 14 |
| Luck | Luck Rating | Continuous | -0.08 to 0.10 |
| Rk | Overall KenPom Ranking | Integer | 1 to 104 |

Table 1. KenPom Feature Data Dictionary

METHODOLOGY

We followed a six-step pipeline from data collection through evaluation:

1. Collected KenPom efficiency data for five seasons (2020–2024).
2. Engineered 7 features from raw ratings and win-loss records.
3. Trained Logistic Regression and Gradient Boosting classifiers for selection, plus a Gradient Boosting regressor for seed prediction.
4. Applied Platt scaling and isotonic regression to calibrate probabilities (fit on training data only).
5. Built conformal prediction intervals for seed forecasts using calibration-set residuals at 90% coverage.
6. Evaluated all models under leave-one-season-out cross-validation (5 folds) to respect temporal structure and avoid data leakage.

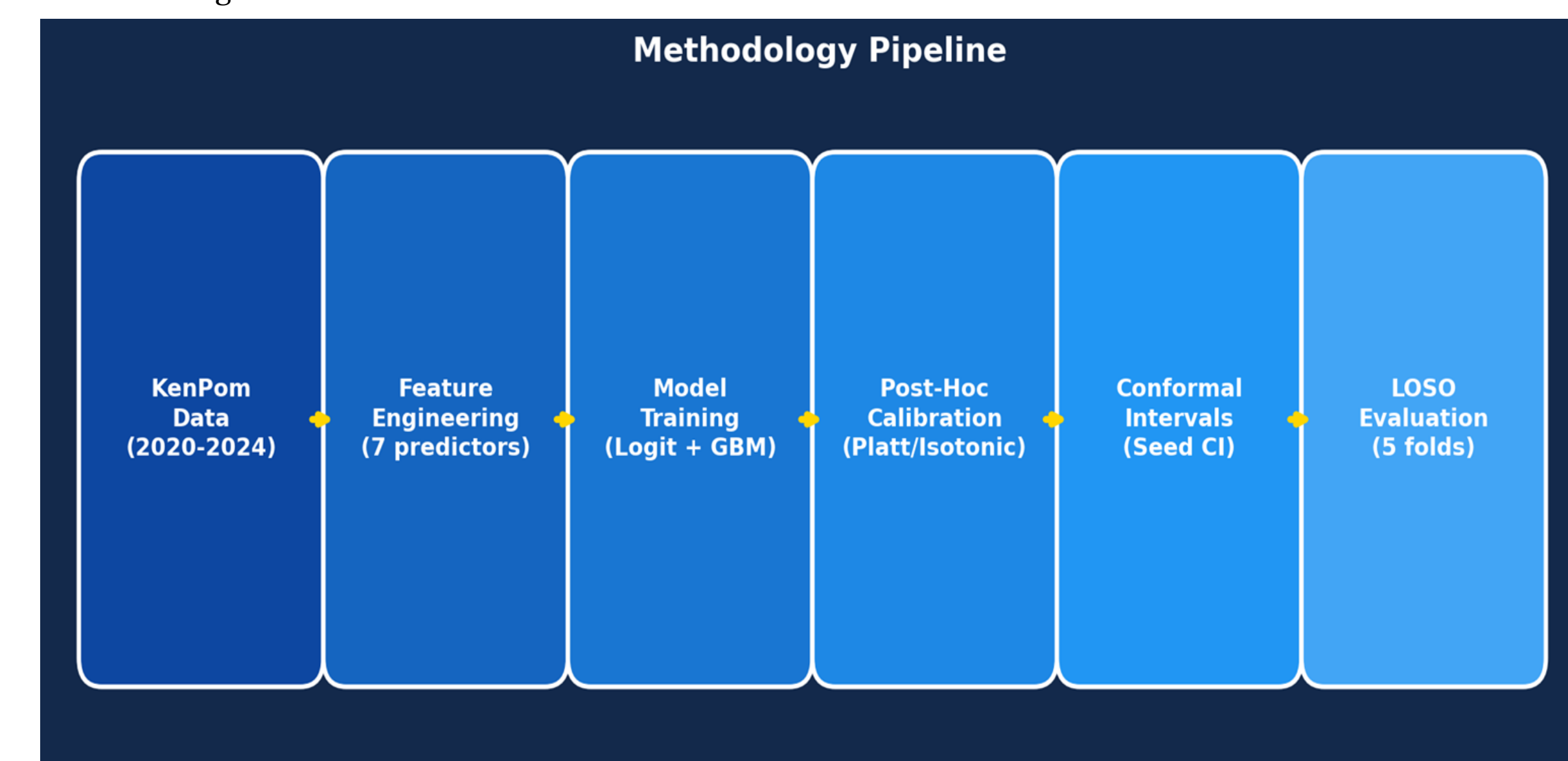


Fig 2. Methodology Pipeline: Data to Calibrated Predictions

MODEL BUILDING & EXPERIMENTAL RESULTS

- We tested six model variants (2 base models × 3 calibration settings) under leave-one-season-out validation. Fig 3 shows train vs. test Brier scores for all variants.
- Logistic Regression with isotonic calibration was the best candidate model, achieving the lowest test Brier score (0.116) and best ECE (0.110). This model was not overfit—the train-test gap was consistent across folds.
- For seed prediction, the GBM regressor achieved a mean absolute error of ~2 seed lines, with conformal intervals covering 89% of true seeds at a mean width of ~10 seed lines.

| Model | Calibration | Brier Score | ECE | Accuracy |
|---------------|-------------|----------------------|----------------------|----------------------|
| Logistic Reg. | None | 0.119 ± 0.053 | 0.125 ± 0.044 | 0.840 ± 0.071 |
| Logistic Reg. | Platt | 0.119 ± 0.045 | 0.135 ± 0.029 | 0.842 ± 0.071 |
| Logistic Reg. | Isotonic | 0.116 ± 0.058 | 0.110 ± 0.057 | 0.846 ± 0.074 |
| GBM | None | 0.126 ± 0.041 | 0.103 ± 0.040 | 0.838 ± 0.050 |
| GBM | Platt | 0.120 ± 0.037 | 0.121 ± 0.021 | 0.832 ± 0.054 |
| GBM | Isotonic | 0.123 ± 0.045 | 0.123 ± 0.036 | 0.840 ± 0.060 |

Green row = Best model (Logistic Reg. + Isotonic Calibration)

Table 2. Model Performance Comparison (LOSO Cross-Validation)

- Isotonic-calibrated Logistic Regression gave the best overall probability accuracy (Brier=0.116, ECE=0.110). The chart shows that after adjustment, the model’s predicted chances line up much better with how often teams actually got selected — especially for bubble teams in the 30–70% range.

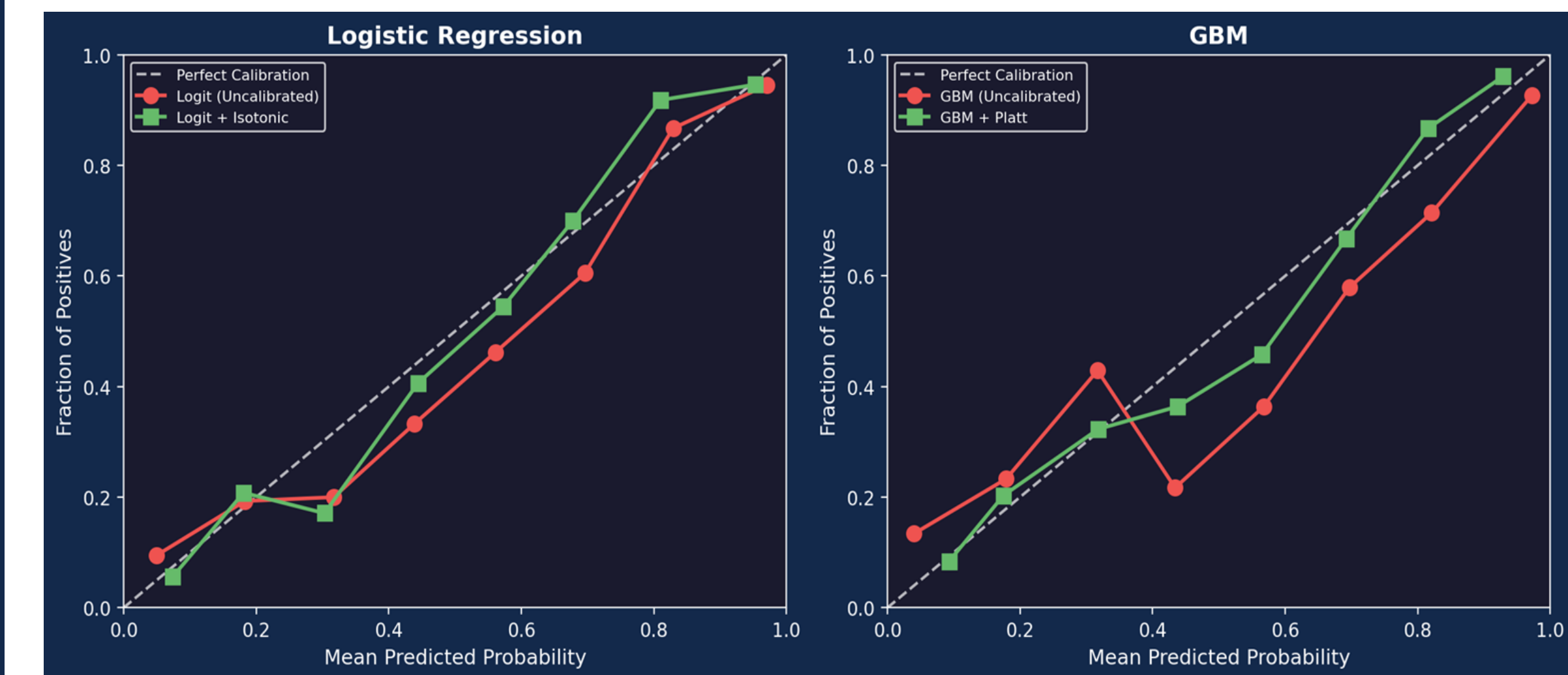


Fig 3. Reliability Diagram: Calibrated vs. Uncalibrated Probabilities

DEPLOYMENT & LIFECYCLE MANAGEMENT

For the NCAA Selection Committee and tournament analysts, this framework provides decision-support in several ways:

- **Improved Decision Confidence:** Instead of a simple yes/no, analysts see the probability each team makes the field. A 45% team is clearly different from a 90% team.
- **Reduced Risk of Errors:** Calibrated probabilities highlight where the committee’s decisions are most uncertain, allowing more focused debate on true bubble teams.
- **Seed Uncertainty:** Prediction intervals give a realistic range of where a team might be seeded, rather than a single guess.
- The model would need annual retraining as each new tournament provides updated selection decisions. Conference realignment (e.g., 2024 changes) may require recalibration.

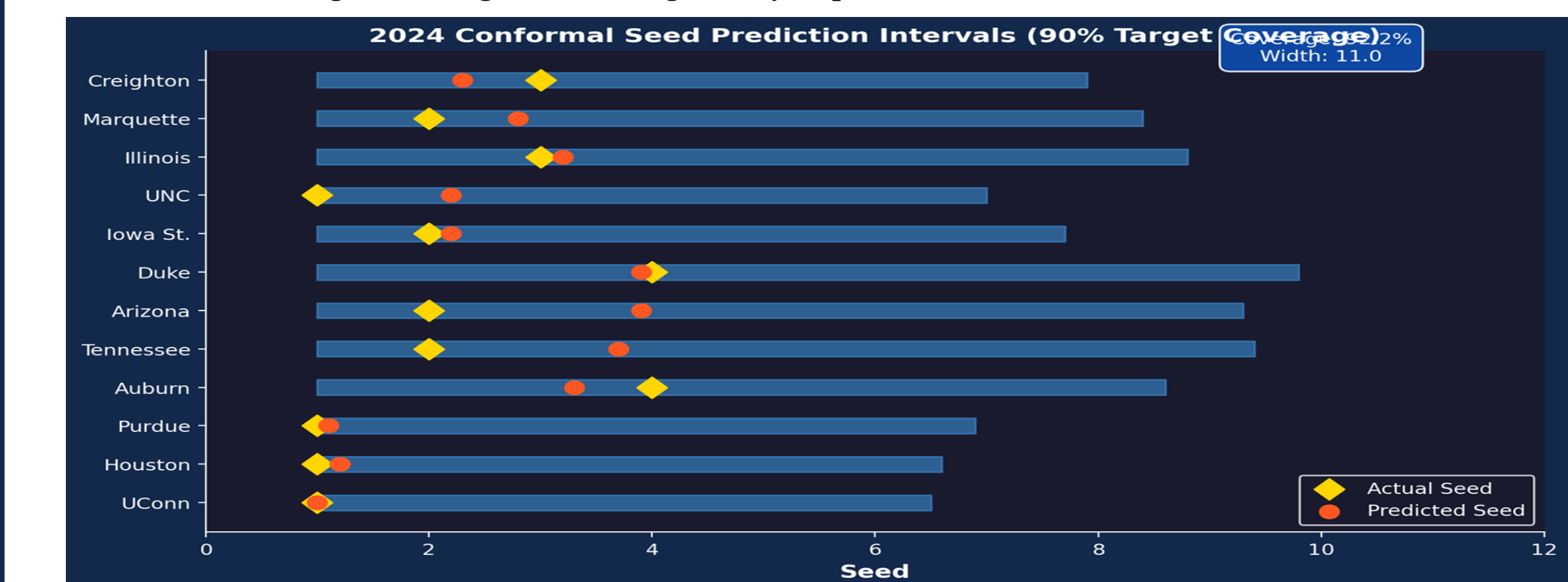


Fig 4. Conformal Seed Prediction Intervals

KEY TAKE-AWAYS

- **Yes—post-hoc calibration does improve selection probability quality. Isotonic calibration reduced Brier score from 0.119 to 0.116 and ECE from 0.125 to 0.110, producing more trustworthy probability estimates for bubble teams.**
- Calibration method choice matters and is model-dependent. Platt scaling helped GBM but did not meaningfully change logistic regression, while isotonic regression improved logistic regression the most.
- Conformal prediction intervals achieved ~89% coverage (target: 90%), providing a practical way to communicate seed uncertainty to decision-makers.
- Of 120 bubble teams (predicted probability 25–75%), only 43% were ultimately selected—confirming this is the zone where calibrated probabilities add the most value over binary predictions.

