Do conference-aware models improve performance and reduce systematic over/under-selection for particular conferences?


Ava Price
asprice@butler.edu


Crystal Longinos
jlonginos@butler.edu


Kate Wombles
kwombles@butler.edu


Maren Berger
mbberger1@butler.edu


Nicholas Kroog
nkroog@butler.edu


Dr. Matthew Lanham (Advisor)
mlanham1@butler.edu

BUTLER
LACY SCHOOL of BUSINESS

ABSTRACT

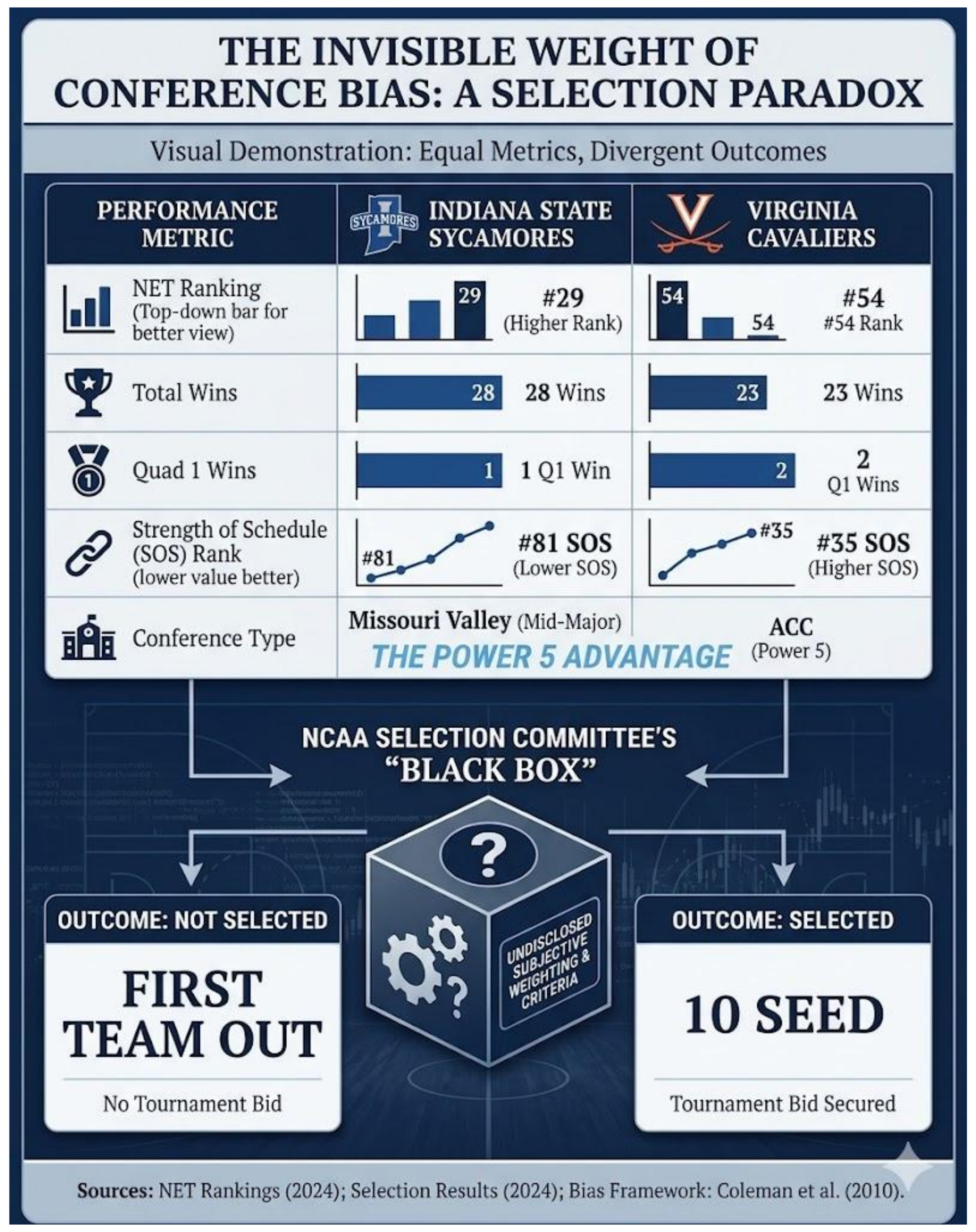
This project examines whether conference aware models improve NCAA selection predictions because they capture real scheduling context or because they reproduce historical bias. Prior studies show conference classification remains predictive even after accounting for team quality. To explore this, we use conference as an explanatory feature and as an auditing dimension. We compare models without conference information to those using one hot encodings, and target encodings. Evaluation combines overall accuracy, conference level calibration, and missed bid disparities to determine whether conference information adds meaningful context, reinforces legacy preferences, or both.

BUSINESS PROBLEM FRAMING

The Challenge:

- **Inconsistency & Subjectivity in Selection** The NCAA Selection Committee faces annual criticism regarding the teams that don't get selected and inconsistent seeding.

Business Problem:
How can the NCAA optimize the selection process to ensure equitable tournament access across all conferences, thereby protecting the integrity of the 'March Madness' brand and maximizing viewership through fair representation?



ANALYTICS PROBLEM FRAMING

We aim to predict NCAA tournament at-large selection using logistic regression.
We built three models:

1. **Baseline model**: excludes conference
2. **One-hot model**: includes conference as dummy variables
3. **Target-encoded model:** includes conference as historical selection rate

We compare models using **accuracy** and **ROC-AUC**.

Finally, we audit predicted probabilities by conference to explore potential bias, asking whether teams from major conferences are more likely to be selected despite similar performance.

DATA

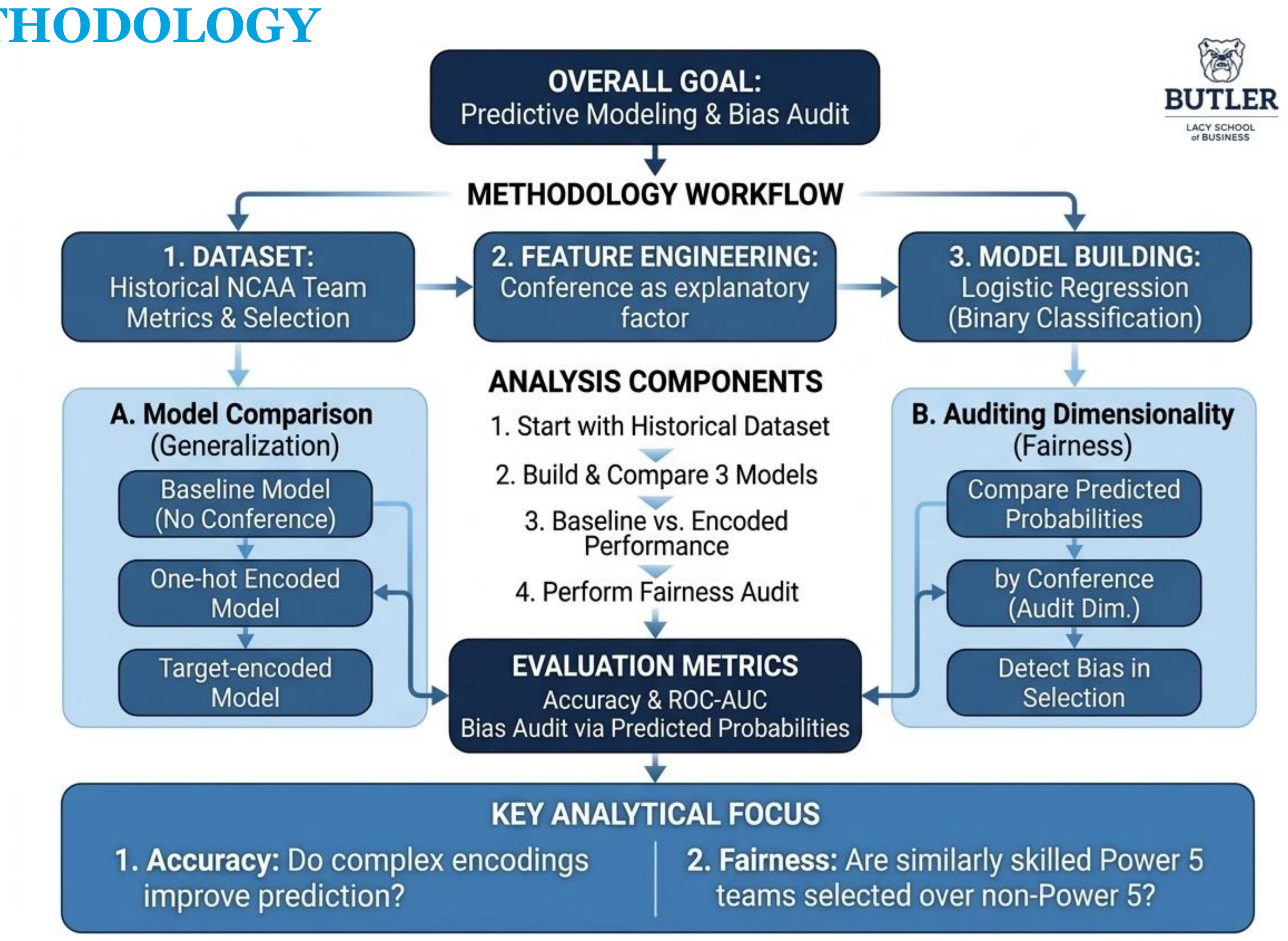
Data Source & Context
We use a historical NCAA dataset containing team performance metrics and tournament selection outcomes.

Key Relationship
Teams with stronger performance (lower NET rank, higher win percentage, and stronger schedules) are more likely to be selected.

Connection to Model
This dataset supports a binary classification task predicting tournament selection (Y = selected). The conference is varied across models (excluded, one-hot, and target encoded) to evaluate whether it improves prediction or introduces bias.

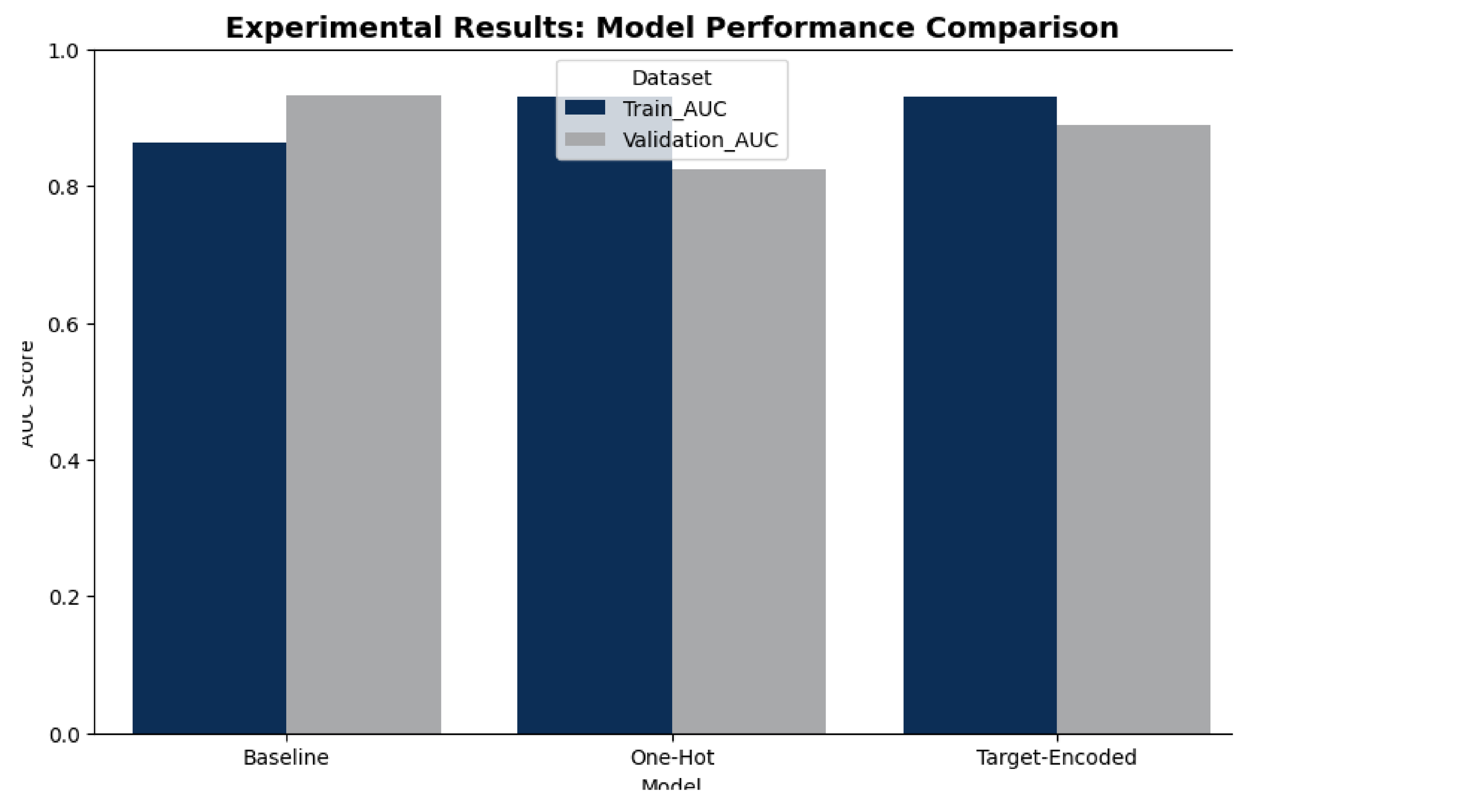
Variable Name	Type	Description	Role in Model
team	Categorical	NCAA team name	Identifier
conference	Categorical	Athletic conference (e.g., ACC, Big Ten, SEC)	Feature (varies by model)
NET_rank	Numeric	NCAA Evaluation Tool ranking (lower = better performance)	Feature
win_pct	Numeric	Overall team win percentage	Feature
Q1_wins	Numeric	Wins against Quadrant 1 opponents	Feature
Q2_wins	Numeric	Wins against Quadrant 2 opponents	Feature
SOS	Numeric	Strength of schedule	Feature
avg_opp_NET	Numeric	Average NET ranking of opponents	Feature
selected	Binary (0/1)	Tournament selection outcome (1 = selected, 0 = not selected)	Target (Y)

METHODOLOGY



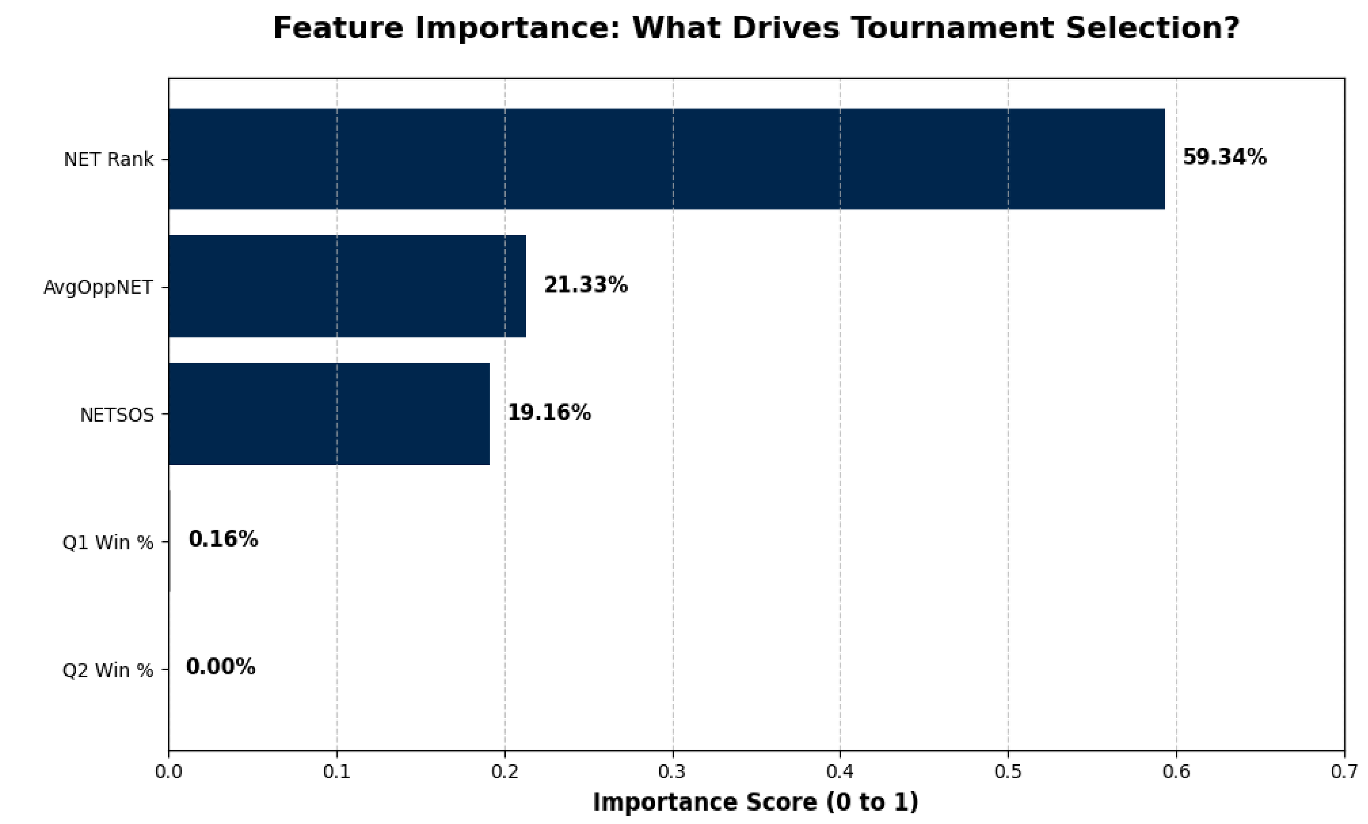
MODEL BUILDING & EXPERIMENTAL RESULTS

Bottom Line: A simpler baseline model generalizes more effectively to new data, whereas complex encoding schemes in this dataset resulted in a performance gap between training and validation.

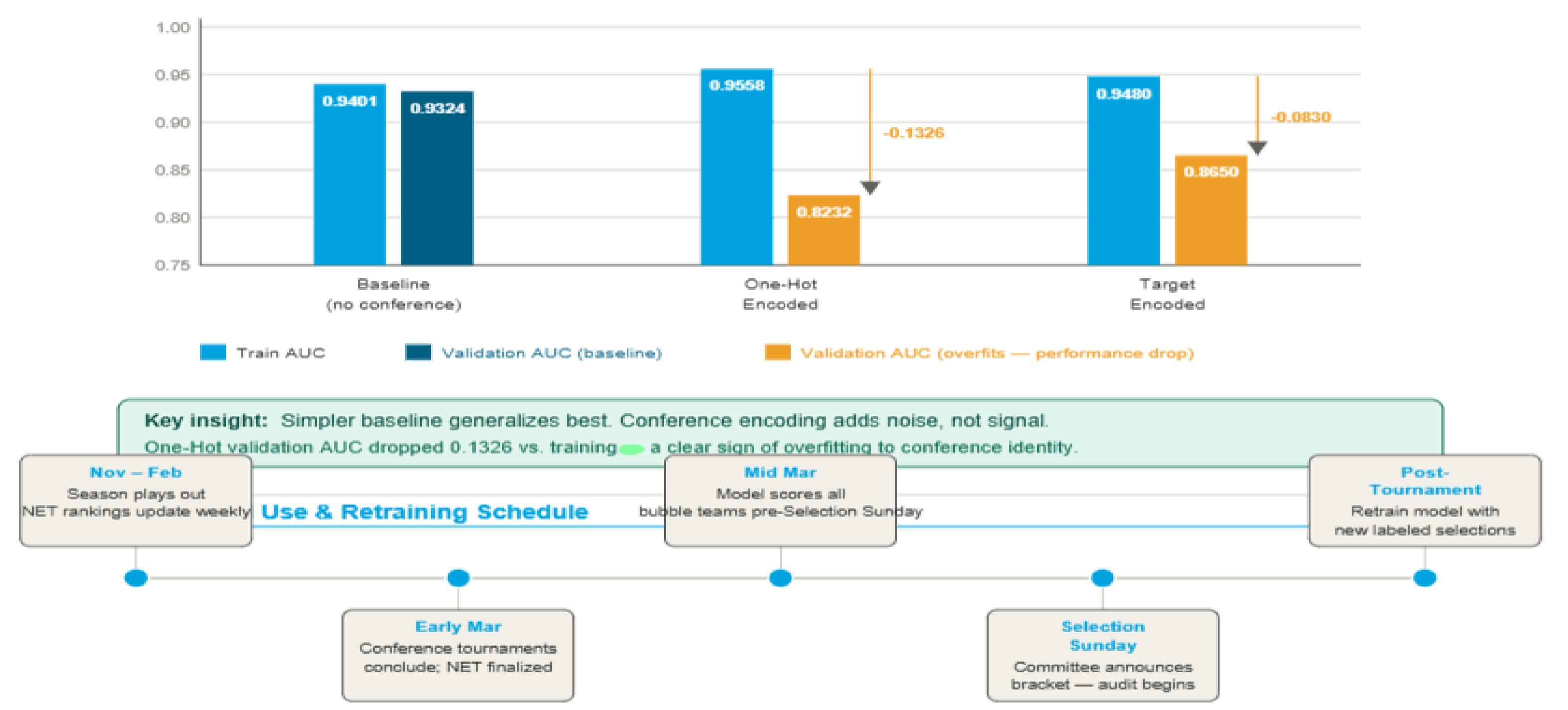


The Goal: To move beyond "black box" predictions and identify which specific metrics the **Random Forest** model prioritizes when selecting tournament teams.

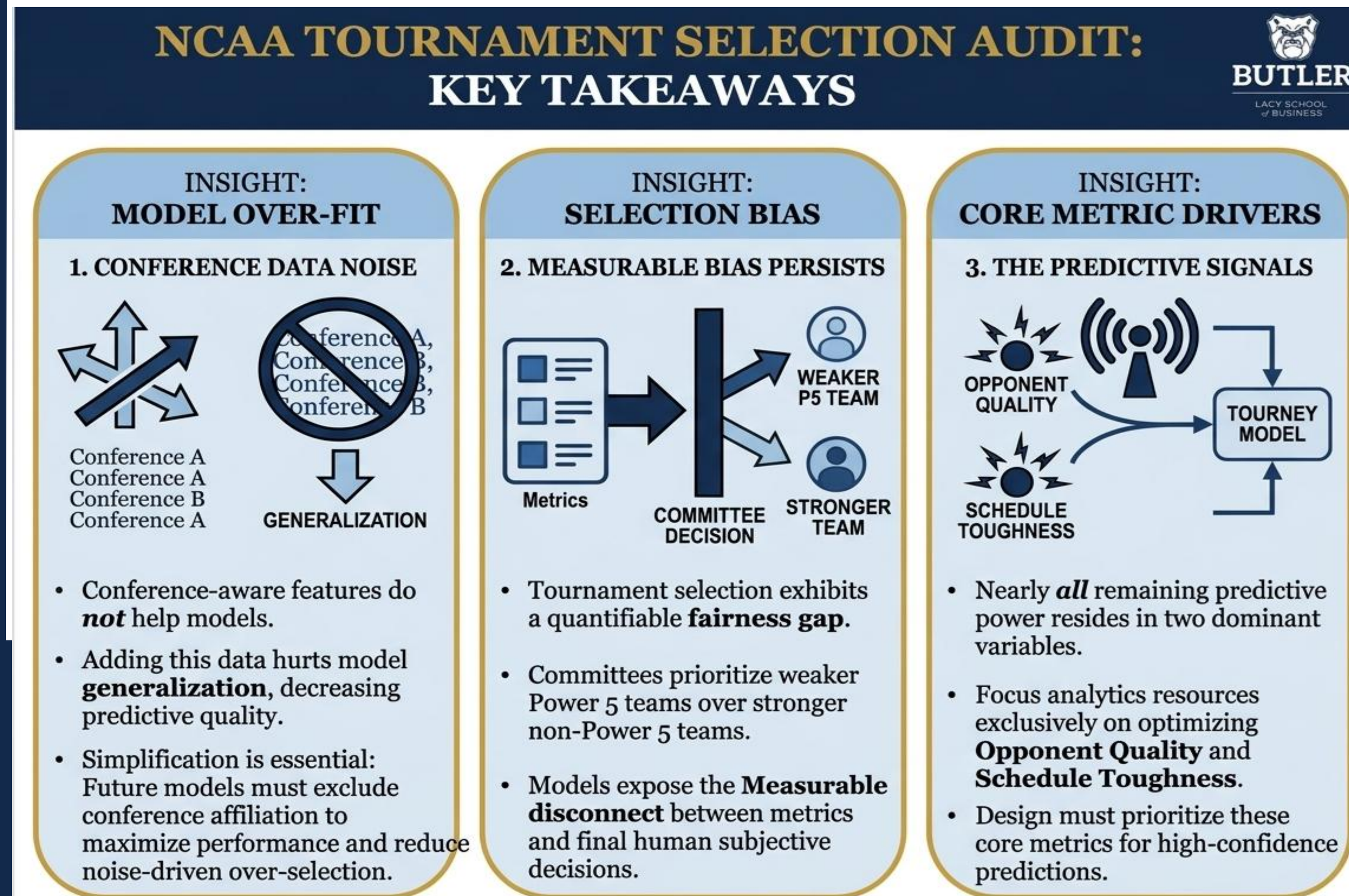
- **Key Findings:**
- **The Power of NET: NET Rank** is the dominant predictor, accounting for **59.3%** of the model's decision-making weight.
- **Strength of Schedule:** Opponent quality (**Avg Opp NET**) and schedule toughness(**NETSOS**) comprise nearly **40%** of the remaining influence.
- **The Quadrant Paradox:** Specific **Q1/Q2 Win Percentages** showed near **0%** importance, suggesting their value is already captured within the overall NET Rank.



DEPLOYMENT & LIFECYCLE MANAGEMENT



Our logistic regression models were validated against historical NCAA selection data. The bias audit adds a second layer of validation: by surfacing cases where statistically stronger teams were passed over for weaker Power 5 programs, the model exposes a measurable gap between data-driven fairness and actual committee decisions.



Personal Development & Outcomes

- DataCamp: Introduction & Intermediate to Python and Intermediate SQL
- Completed Machine Learning Using SAS® Viya course
- Developed skills using Google Colab and Gemini for coding & data analysis
- Developed presentation skills at Butler's URC

sas CERTIFIED
ADVANCED ANALYTICS LEARN sas
NCAA MEN'S FINAL FOUR ANALYTICS CHALLENGE