Predicting the Sweet 16: AN XGBoost Model for March Madness Bracketology

Mena Tobia
Kyle Yusong Yao
Andrew Galas
Noa Druyan

UNC SCHOOL OF DATA SCIENCE AND SOCIETY

The Problem

- Predict if a team can make the top 16 for the postseason based on pass statistics
  - Our project focuses on using team statistics to predict whether a college basketball team will make it to the top 16 (Sweet Sixteen) for the postseason. We will be using performance metrics like shooting efficiency, rebounds, turnovers, defensive stats, and more, to associate the outcome of a game.

Data Collection

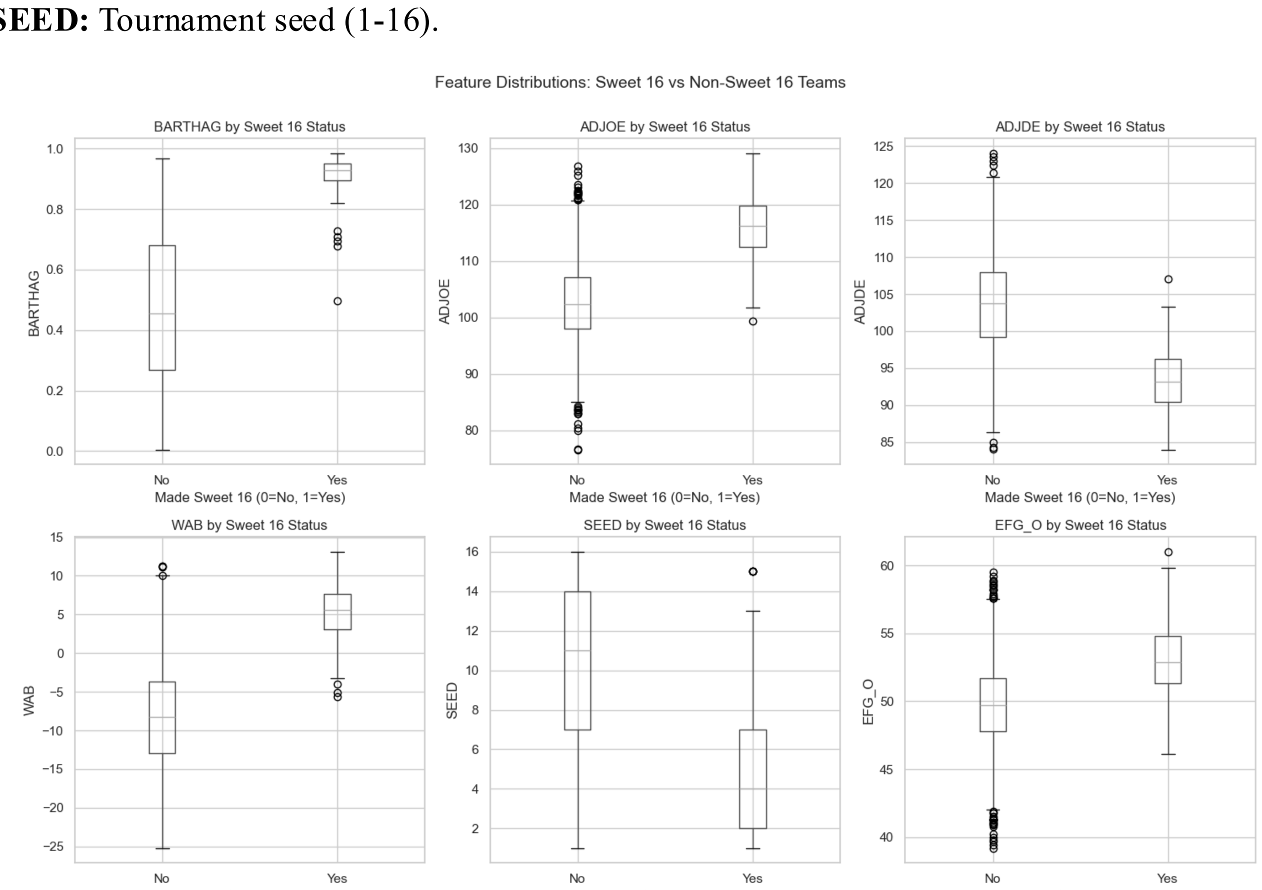
- Our data set if from http://barttorvik.com/trank.php
- This dataset is one table with 3525 samples (rows) and 23 features (columns)
- Each row represents a different team in a given season. Features include many different team statistics including offensive and defensive statistics.
- Due to the data only including Division 1 teams, our conclusions can't be generalized to other levels, like other Divisions of College sports or professional leagues.
- The data also can not capture other factors such as team chemistry, injuries, or coaching strategy.

Data Preparation

- **Temporal Train/Test Split**: instead of a random split, we split based on time
  - **Training**: 2013 -2022 data (3160 teams)
  - **Testing**: 2023 data (363 teams)
- **Leakage Prevention**: We dropped the G (Games Played) and W (Wins) columns because knowing how many games a team played reveals how far they went in the tournament, which is what we are trying to predict
- Preprocessing:
  - For **Logistic Regression**, we standardized numerical features (scaled to mean=0,std=1) because linear models are sensitive to the magnitude of numbers.
  - For **XGBoost**, we did not scale features because tree-based models are scale-invariant.

Data Exploration

- **Target Variable Imbalance:** We identified a big class imbalance. Only 4.5% of teams (160 out of 3,523) actually make the Sweet 16. If we didn't address this a model could just predict "No" every time and still get 95% accuracy .
- **Key Features Analyzed:** We focused on 20 features, but the analysis highlighted three major ones.
  - **BARTHAG:** Power rating (Estimate of a team's chance of beating an average D1 team).
  - **WAB:** Wins Above Bubble (How many more wins a team has compared to average "bubble" team).
  - **SEED:** Tournament seed (1-16).



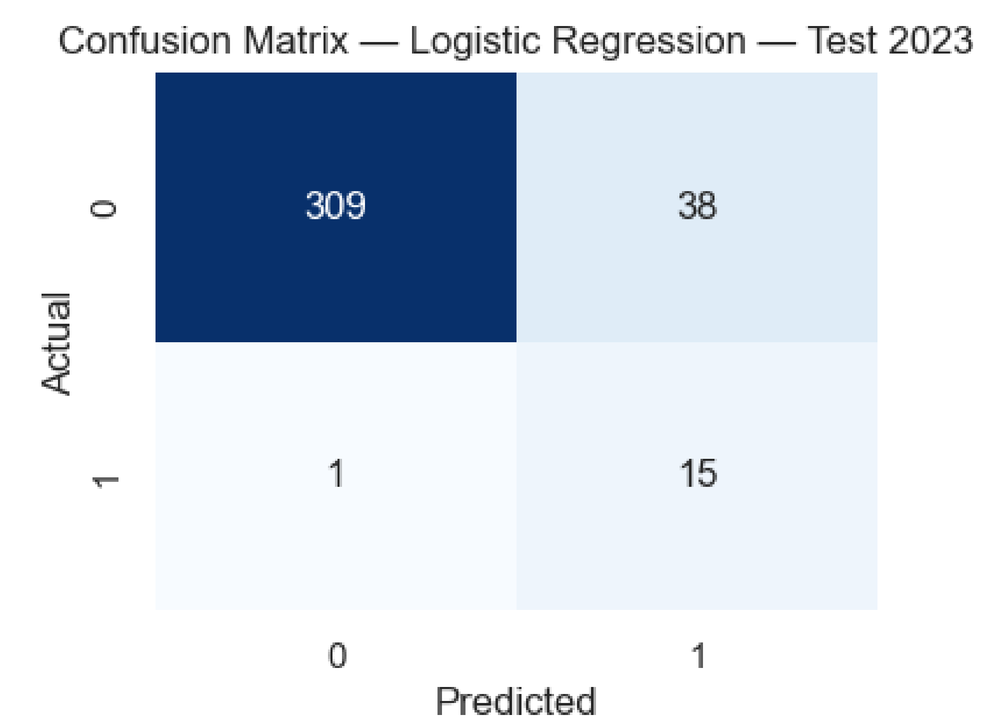
Model Building

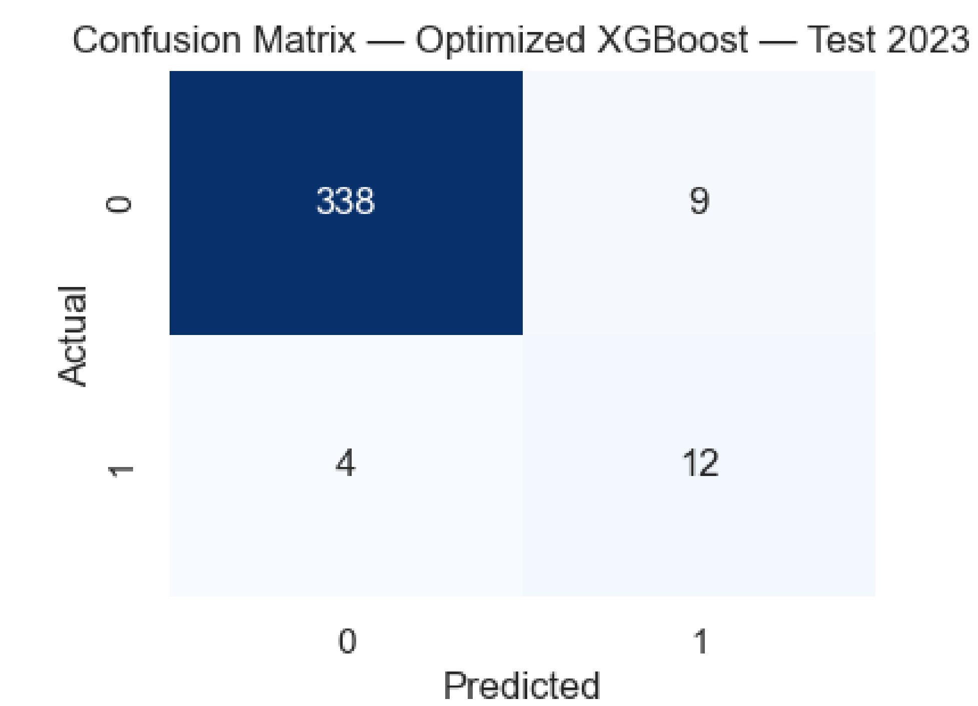
Model 1: Logistic Regression (Baseline):

- *Why:* To establish a benchmark using a simple, interpretable model.
- *Configuration:* Used class_weight='balanced' to force the model to pay attention to the rare "Sweet 16" teams.

Model 2: XGBoost (Champion Model):

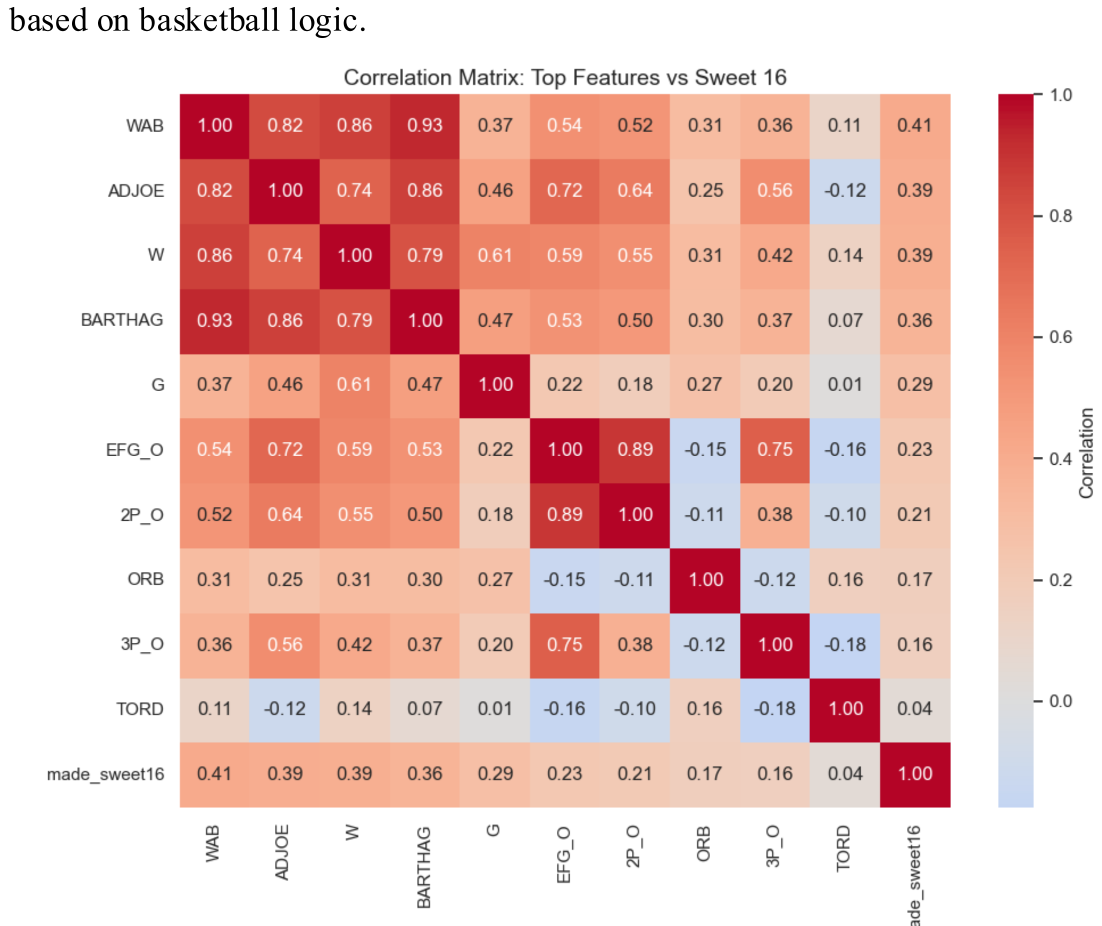
- *Why:* It captures non-linear relationships better than logistic regression and handles complex interactions between stats.
- *Optimization:* We tuned the scale_pos_weight parameter. We found that a multiplier of **3.0x** was optimal. This told the model that missing a Sweet 16 team was 3 times worse than a false alarm. This boosted our Recall from 68.8% to 75.0%.

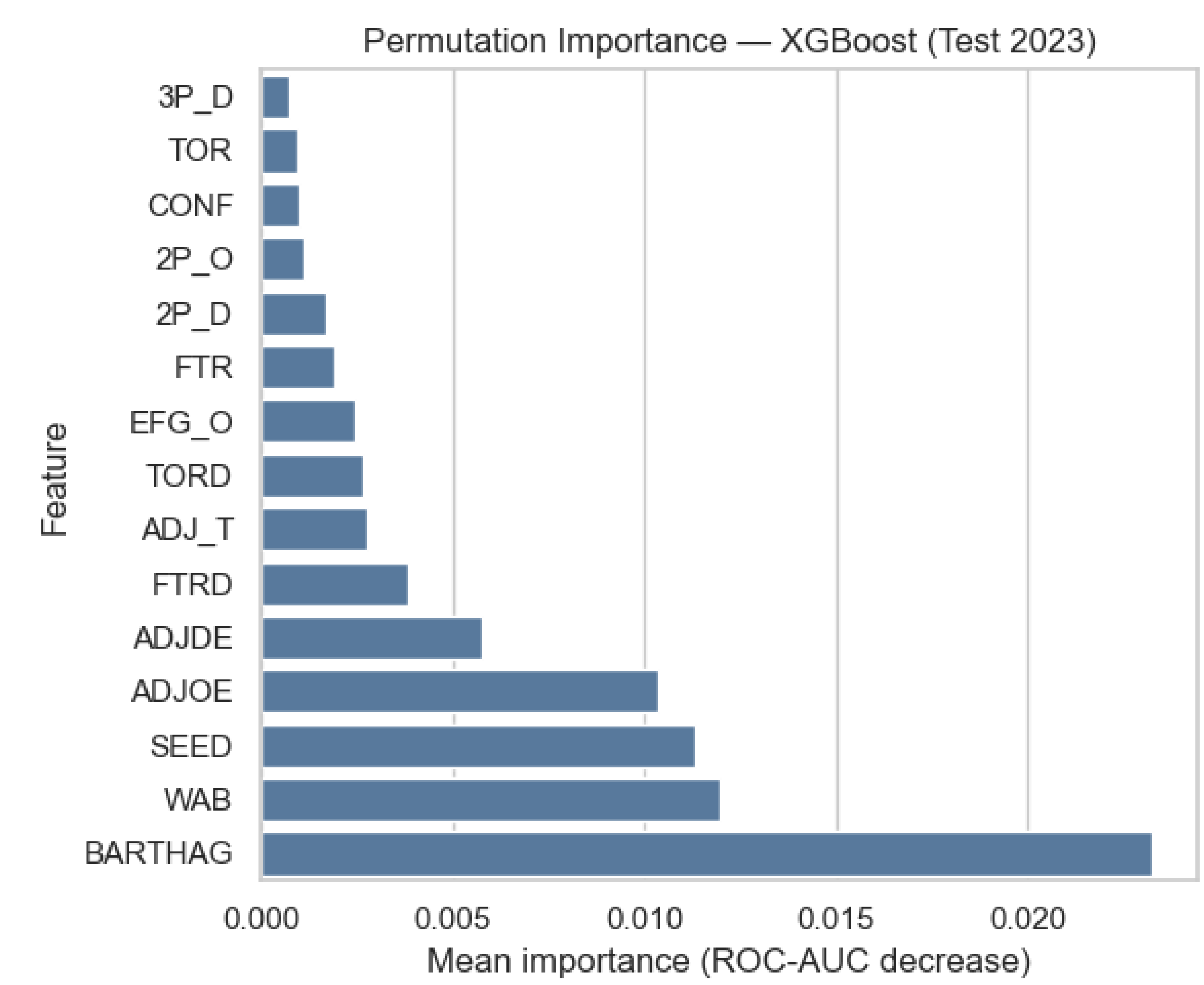




Model Evaluation

- **The Winner:** XGBoost outperformed Logistic Regression significantly.
  - **Accuracy:** 96.4% (XGBoost) vs 89.3% (LogReg).
  - **ROC-AUC:** 0.981 (XGBoost) vs 0.960 (LogReg)
- **Interpretation of Success (Recall vs. Precision):**
  - Our goal was to **catch** the Sweet 16 teams (high Recall).
  - The optimized XGBoost model correctly identified **12 out of 16** actual Sweet 16 teams in the 2023 test set (75% Recall).
  - *Trade-off:* To catch those teams, the model generated 9 "false alarms" (teams predicted to make it who didn't). In bracketology, this is acceptable because we prefer to be optimistic about strong teams than miss a Cinderella story.
- **Feature Importance (Validation):** The model identified **BARTHAG** (Power Rating) as the #1 most predictive feature, followed by **WAB** and **SEED**. This validates the model "makes sense" based on basketball logic.





Model Deployment

- **Use Case:** This model could be deployed as a "Bracket Helper" tool for Selection Sunday, flagging teams with high probabilities of deep runs.
- **Risks:**
  - **Drift:** The model was trained on 2013-2022. Changes in NCAA rules (like the transfer portal or 3-point line distance) might make older data less relevant.
  - **Uncertainty:** Small sample size means one or two injuries could swing the results significantly.