NCAA March Madness Tournament Prediction

Project Goal:

To build a predictive model for NCAA Men's March Madness tournament games using historical performance data, with the goal of forecasting outcomes and uncovering key factors that drive wins in high-variance, single-elimination matchups.

Why It Matters:

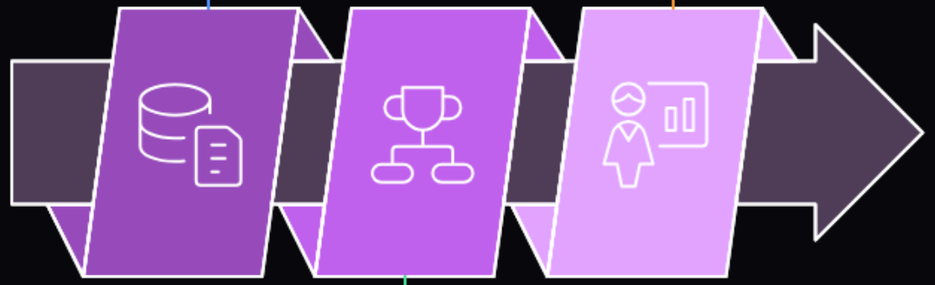
The tournament's structure (63 games, 68 teams, 9.2 quintillion possible bracket combinations) creates a unique modeling challenge. Even domain experts struggle to beat simple heuristics like picking higher seeds. This project explores whether machine learning can outperform human heuristics and improve prediction quality.

Data Collection

- Gathered NCAA team, game, and conference data across multiple seasons
- Combined historical regular season box scores with tournament results
- Restructured raw game logs into consistent team-vs-opponent format
- Calculated base metrics (e.g., FGM/FGA, possessions, efficiencies)
- Integrated advanced stats such as KenPom ratings and Massey Ordinals
- Aggregated statistics at multiple levels: game, season, team, and opponent

Modeling & Evaluation

- Handled categorical variables with one-hot encoding
- Explored power conference vs. mid-major segmentation
- Developed baseline heuristic (seed-based prediction)
- Trained and evaluated logistic regression and XGBoost models
- Conducted robust model evaluation: AUC, precision, recall, F1, confusion matrix



Modeling & Evaluation

- Mapped team-level stats onto tournament bracket matchups
- Engineered matchup-specific features (e.g., seed gap, KenPom delta)
- Labeled outcomes for binary classification (win/loss per matchup)

Overview

Data Sources

- Historical NCAA tournament results
- Regular season performance data
- Advanced stats from KenPom and Massey Ordinals
- Custom-engineered features from box scores

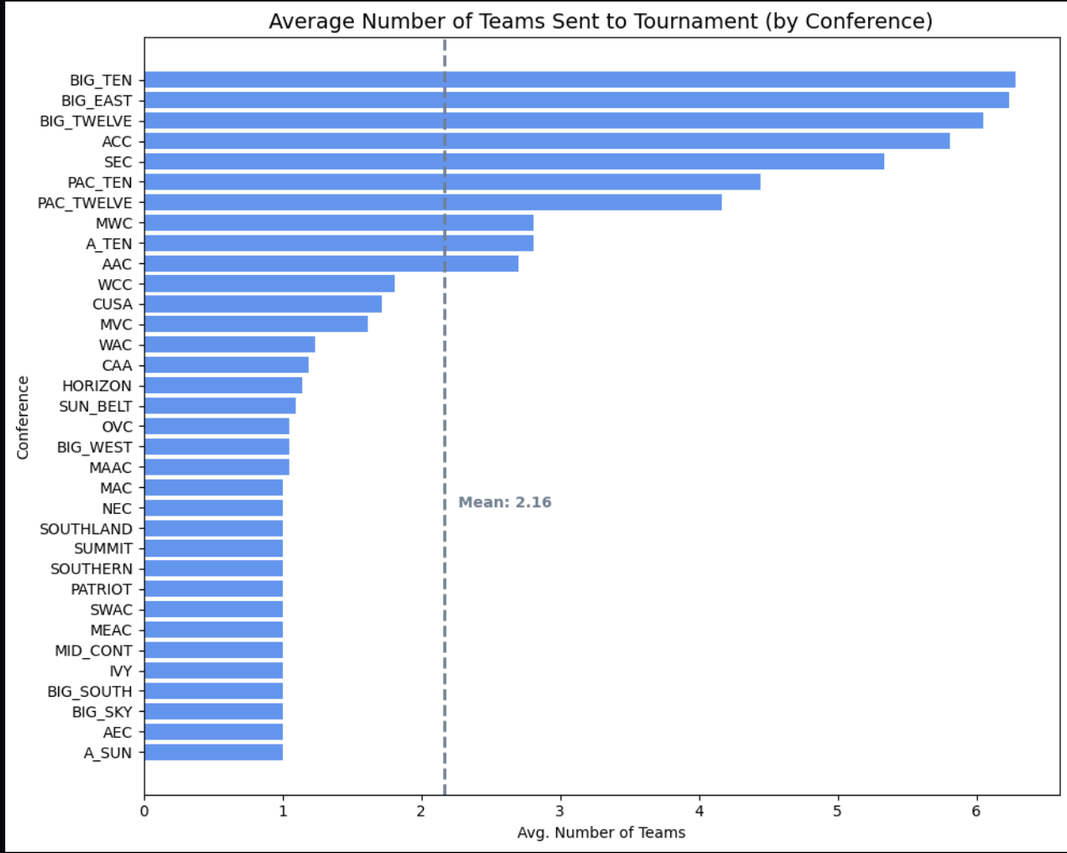
Preprocessing Highlights

- Transformed game-level data to consistent team-vs-opponent structure
- Features engineering including:
  - Per-game and aggregate stats
  - Delta features (team - opponent)
  - Opponent-adjusted performance metrics
  - Ranking and seed-based context

Conferences Aren't All Built Equally...

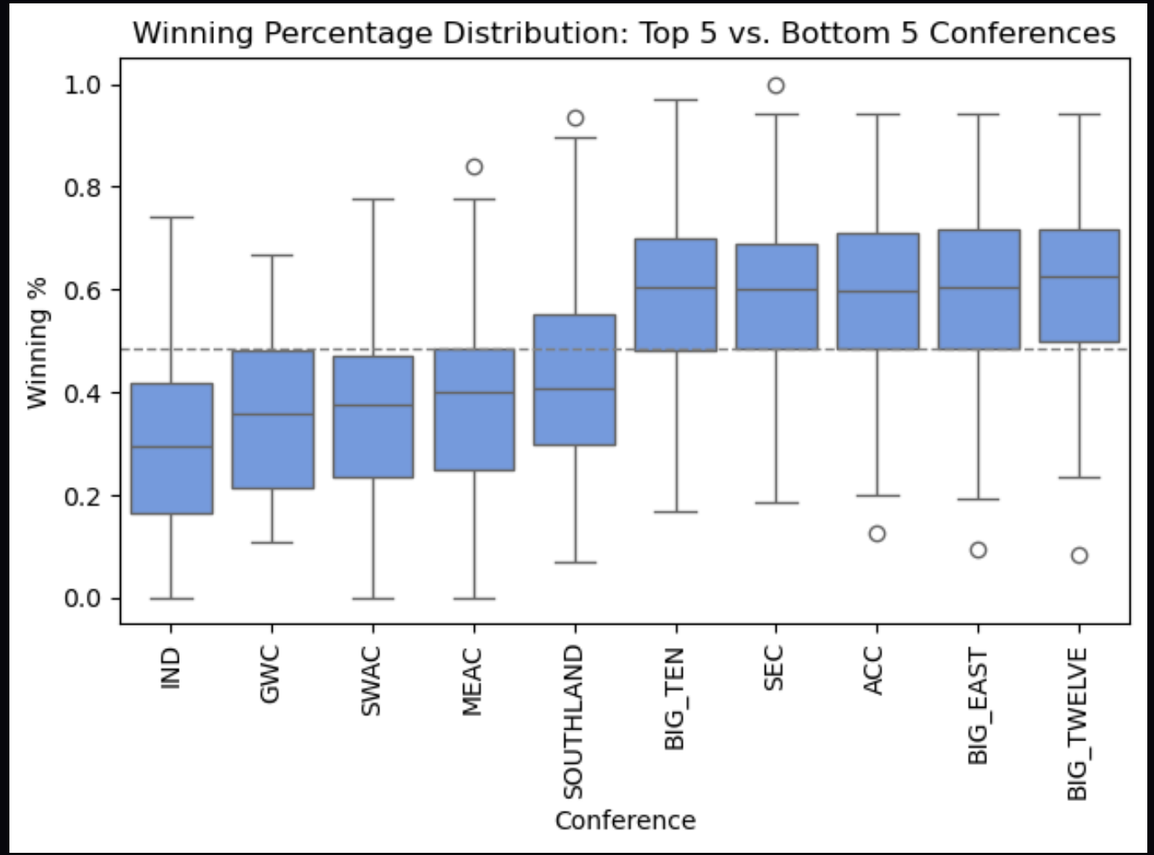
Although the dataset is balanced in terms of class (win/loss), the makeup of teams and their conferences indicates an inherent imbalance.

Though the number of teams entering the tournament from each conference varies from one season to another, the **"Power Conferences"** consistently contribute a significantly higher proportion of tournament teams.



As you can see, the same "Power Conference" teams also boast the highest historical average winning percentages, while the mid-major conferences dominate the list of lowest regular season winning percent.

The programs in the larger conferences tend to have the largest athletic budgets but are also generating large sums of revenue relative to other conferences through media and network contracts.



Modeling & Preliminary Results

Rule-Based Baseline

To ground our model evaluation to a meaningful reference point, we constructed a baseline predictor that mirrors a common heuristics-based approach:

- Pick the higher seed to win each matchup
- In cases where both teams have the same seed, use final KenPomRank as a tiebreaker
- Baseline accuracy of 70.22% highlights the predictive value captured by seeding and advanced team metrics like KenPom rankings

Logistic Regression

- Overall test accuracy of ~71.6% outperforms the naive "chalk" baseline
- Top predictors included score differential, possession efficiency, and opponent-adjusted rebounding and foul metrics
- Model highlighted the predictive power of relative strength features (delta stats) over raw performance metrics
- Strong alignment between training and test performance suggests our feature set and model complexity are not over or underfitting

XGBoost

- Overall test accuracy of ~71.2%
- Most important features included KenPom ranking deltas, seed gaps, and efficiency differentials between teams
- The model showed stronger signal from features that combined team context, strength-of-schedule adjustments, and possession-based efficiency (non-linear relationships)

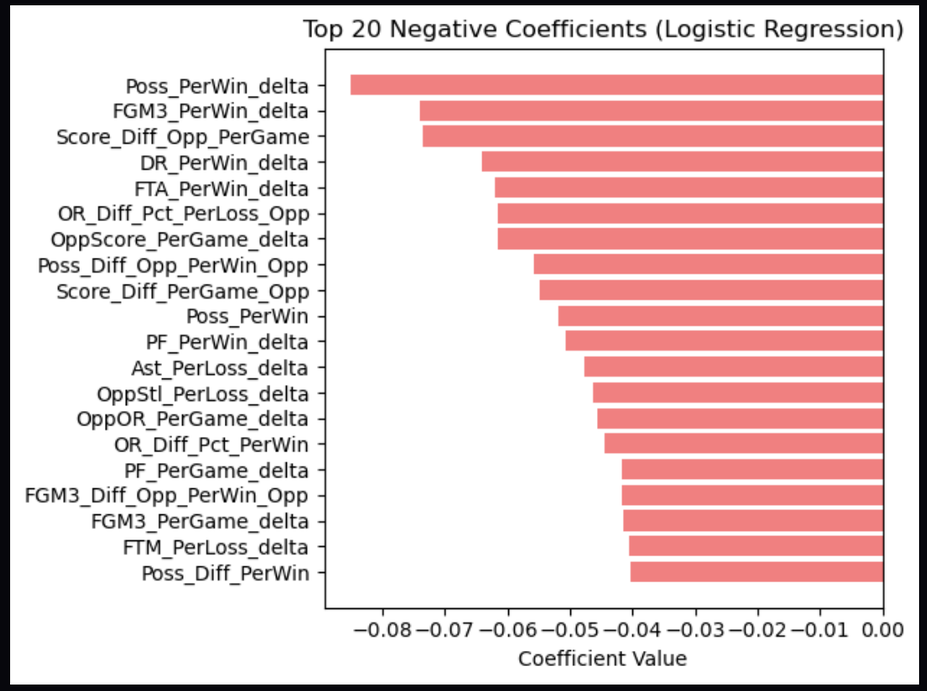
Further Model Iterations

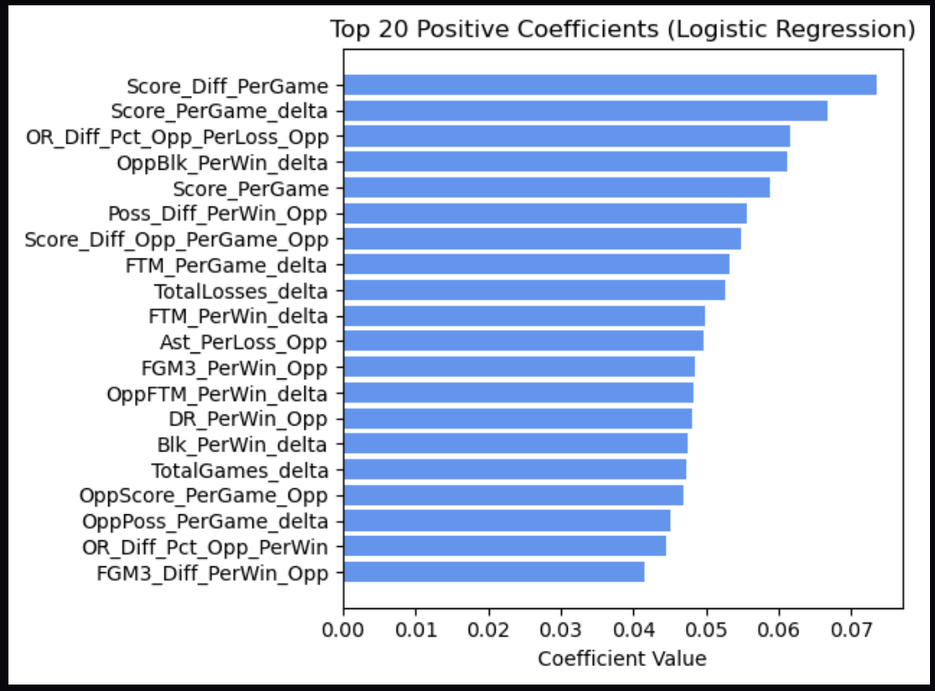
- Develop ensemble techniques to combine strengths of linear (logistic) and nonlinear (XGBoost) models for improved predictive performance
- Incorporate additional advanced metrics
- Rigorous feature selection to see if there are features in the current model that are more noise than informative to the model(s)

Logistic Regression Coefficients

Logistic Regression

- Score-based metrics such as emerge as the strongest positive predictors and carry significant predictive weight
- Opponent rebounding and rim protection features like show that limiting opponents' second-chance points and minimizing shot-block disruption are subtle but meaningful contributors to win probability
- Delta features indicate that the difference between a team and its opponent across advanced stats like FTM, BLK, and FGM3 are also important predictors

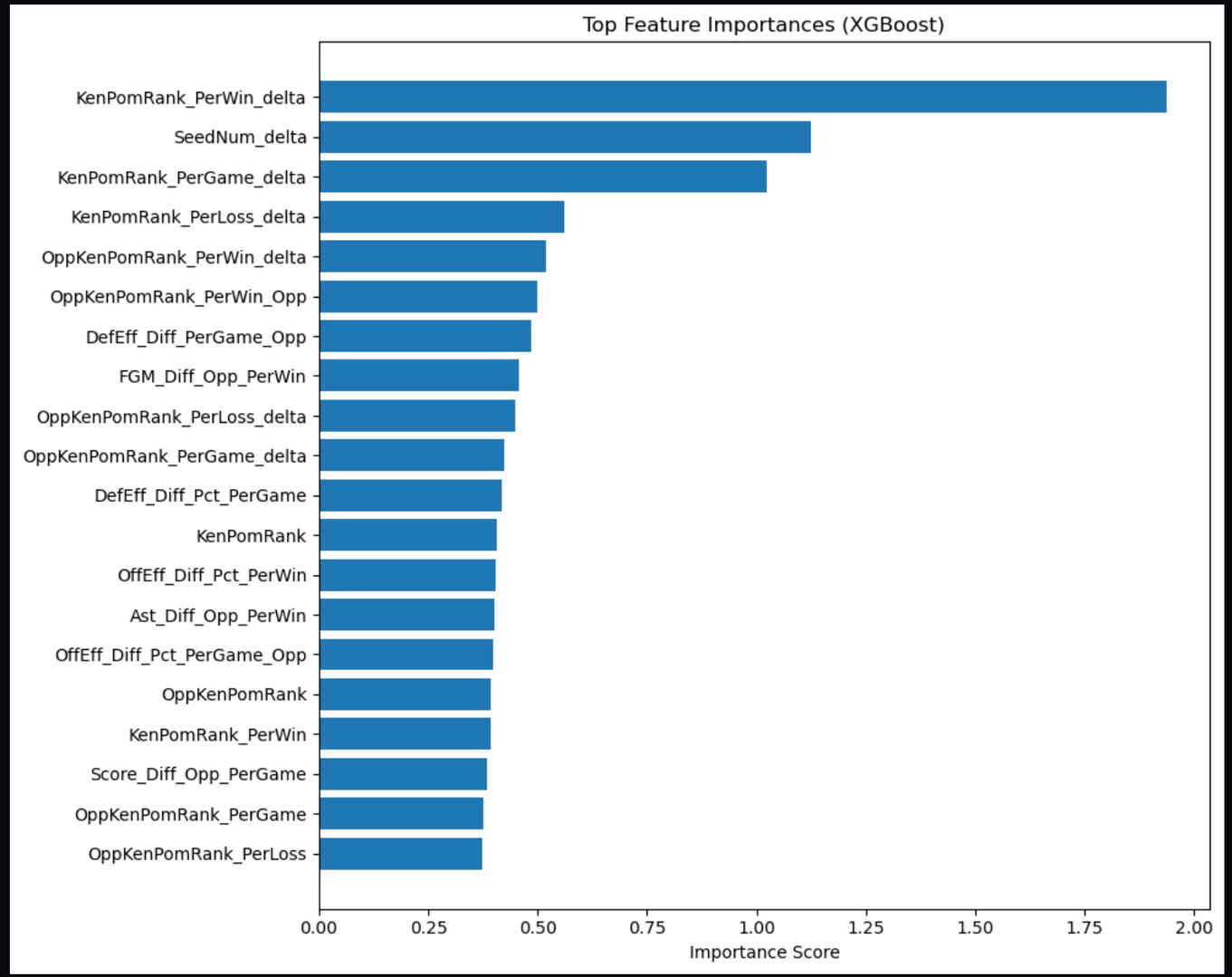




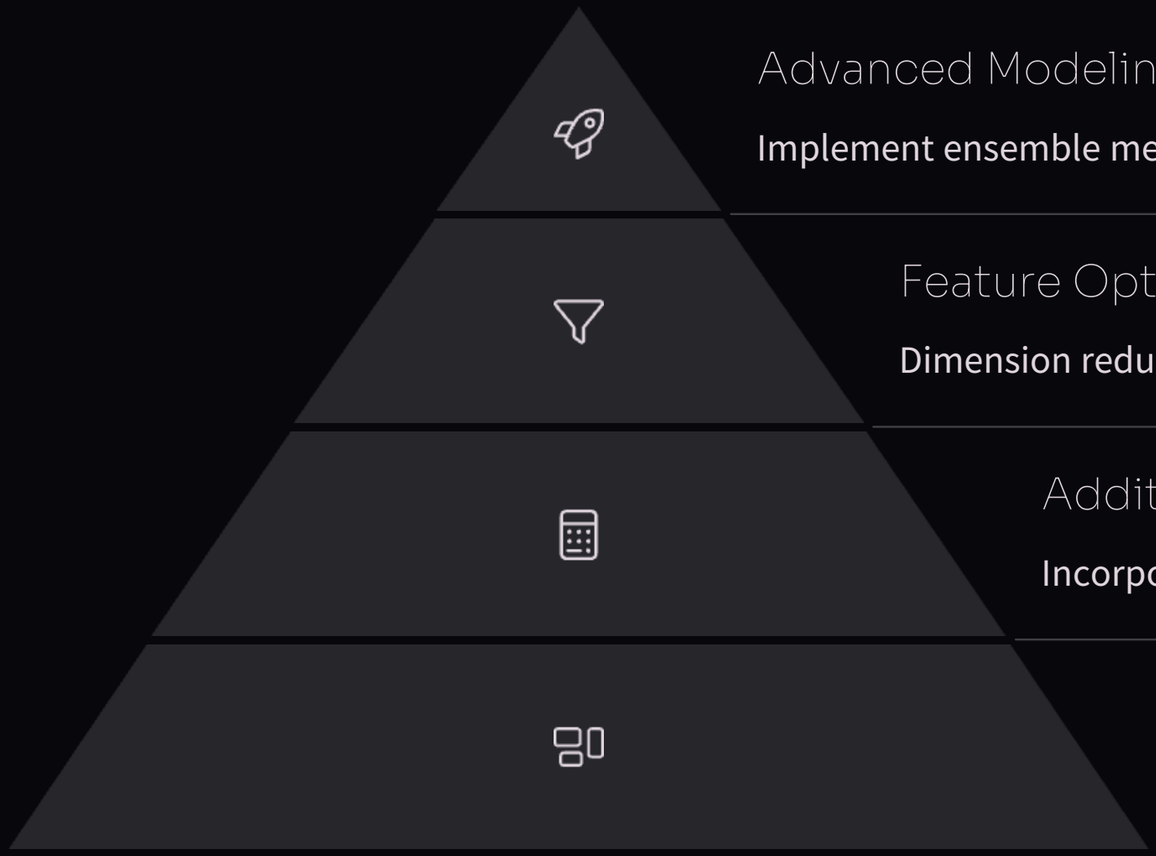
XGBoost Feature Importances

XGBoost

- The number of KenPom related statistics signals that the model is getting the most information from the advanced / non-linear metrics in the data
- Differential / delta features seem to have the highest magnitude impacts on the model
- Adjusted-Efficiency Metrics are informative to our XGB model while less impactful in our logistic model



Next Steps



Advanced Modeling

Implement ensemble methods

Feature Optimization

Dimension reduction and rigorous feature selection

Additional Metrics

Incorporate more advanced basketball analytics

Simulation Framework

Build bracket simulation for tournament-wide predictions

The next phase will focus on refining the model through dimension reduction and feature selection to improve performance. We'll add more advanced basketball metrics and develop a comprehensive bracket simulation framework that can generate full tournament predictions with confidence intervals.