NCAA March Madness Predictions

A Data Science Approach to Modeling Tournament Outcomes

Apr 2025

The Challenge of Predicting March Madness

- Single-elimination format, 68 teams, 63 games
- **9.2 quintillion** possible brackets
- Even experts struggle — picking **higher seeds yields ~70% accuracy**

Project Goal

- Build **predictive model** to forecast NCAA tournament games
- Move **beyond basic heuristics** like seeding
- Uncover **key factors** driving success in high-variance matchups

Data & Preprocessing

Data Sources

- Regular season and tournament results
- Advanced metrics (KenPom, Massey Ordinals)
- Custom features: scoring margins, tempo, efficiency

1. **Restructured box-score data to model interpretable format**
2. **Engineered matchup-specific features** (e.g., KenPom deltas, seed gap)
3. **Aggregated stats at team-season level**
4. **Calculated advanced metrics and features**

Process Overview

Data Preprocessing

Load / preprocess regular season and tournament data

Exploratory Data analysis

Explore different features of the dataset

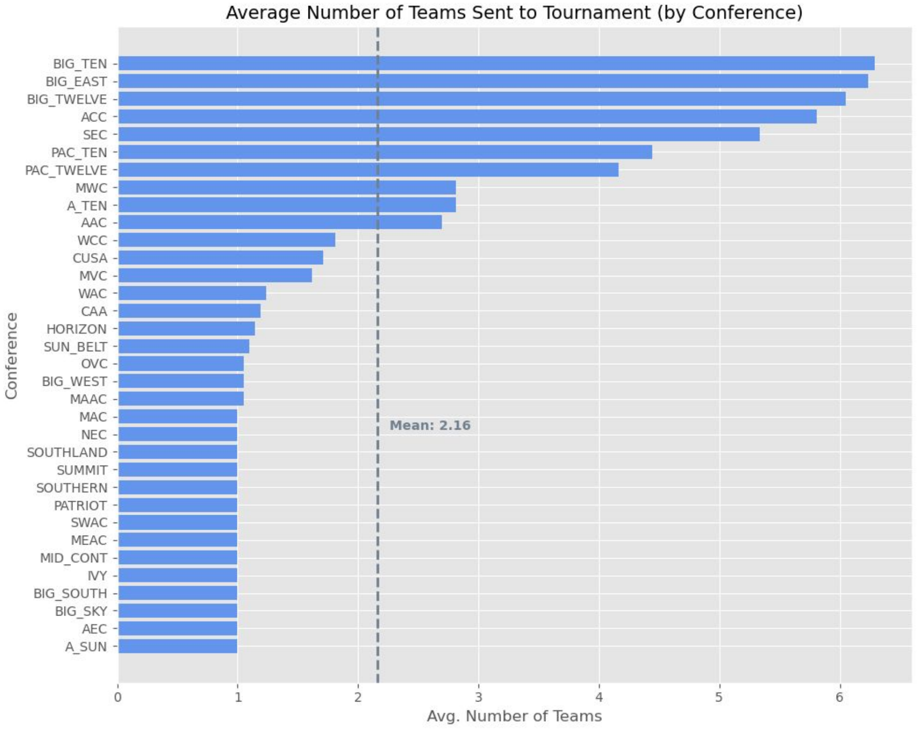
Model Preparation

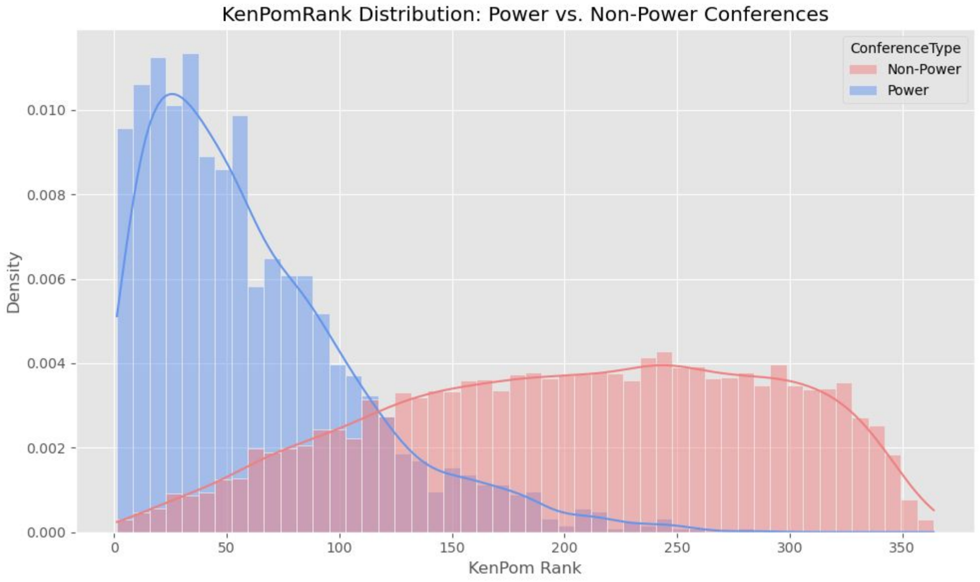
Engineer new features, encode categorical variables, and prepare for modeling

Modeling & Evaluation

Predict outcomes using ML / statistical models, and evaluate results

Conferences & Competition





Power vs. Non-Power Conferences

- Power conferences (**Big Ten**, **SEC**, **Big 12**, etc.) dominate tournament participation
- These leagues consistently field stronger teams confirmed by:
  - Higher average KenPom rankings
  - Tougher strength of schedule
  - Higher avg. tournament bid counts

Model Comparison

Logistic Regression

- Accuracy: 71.5%
- AUC: 0.795
- Highlights:
  - Stronger than naive seed-based baseline (~69.7%)
  - Top features: Seed gap, KenPom deltas, opponent strength

XGBoost

- Accuracy: 71.1%
- AUC: 0.798
- Highlights:
  - Better handling of non-linear feature interactions
  - Heavy emphasis on KenPom ranking deltas and matchup-specific metrics

Conclusion

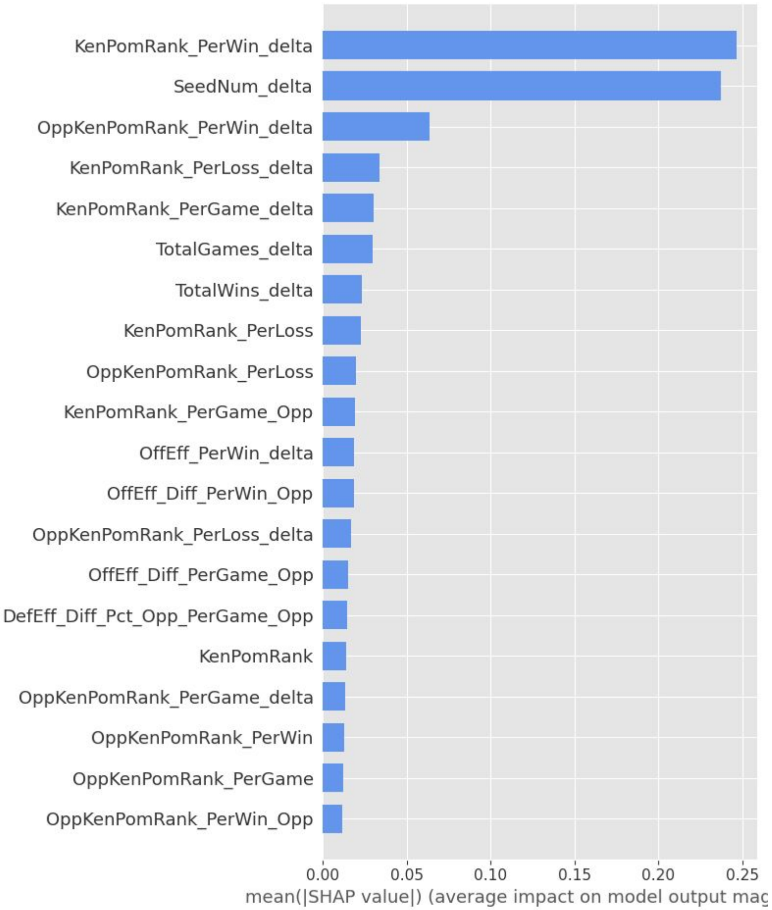
Both models outperform baseline heuristics, with XGBoost offering marginal improvements and deeper signal extraction.

Model Interpretability

What Drives Predictions?

Based on the **SHAP** analysis:

- **KenPom dominates:** most top SHAP values are matchup strength differentials
- **Seed differentials** are highly predictive (upset are upsets for a reason)
- **Opponent quality** features were helpful in determining wins/losses



Impact & Next steps

Based on the results of our models:

- Validated approach > baseline heuristic
- Model understands strength-of-schedule, efficiency, and matchup dynamics

1. Add ensemble models for robustness
2. Look into unsupervised learning techniques
3. Build an interactive dashboard