

Modeling the Madness

Predicting NCAA Men's Basketball Tournament Outcomes

Samira Diabi Genna Moellering Amelia Janaskie

Setting the Stage

Purpose: Predict March Madness Tournament Outcomes

The Odds: 1 in 9,223,372,036,854,775,808

- Used data on game-level statistics
 - Years 2003 2023: Tournament, Regular Season, Tourney Seeds, Team Coaches, and Team Conferences
- Link

The Game Plan



Which teams will make it into the NCAA Men's Basketball Tournament in 2023?

Which teams will win each round of the tournament bracket?

Gaussian Naïve Bayes

Logistic Regression

Random Forest Classifier

XGBoost Classifier

Have tournament outcomes become harder to predict in the Name Image & Likeness (NIL) Era where NCAA athletes can be paid by universities?

Gaussian Naïve Bayes

Logistic Regression

Random Forest Classifier

Support Vector Machine Classifier (SVM)

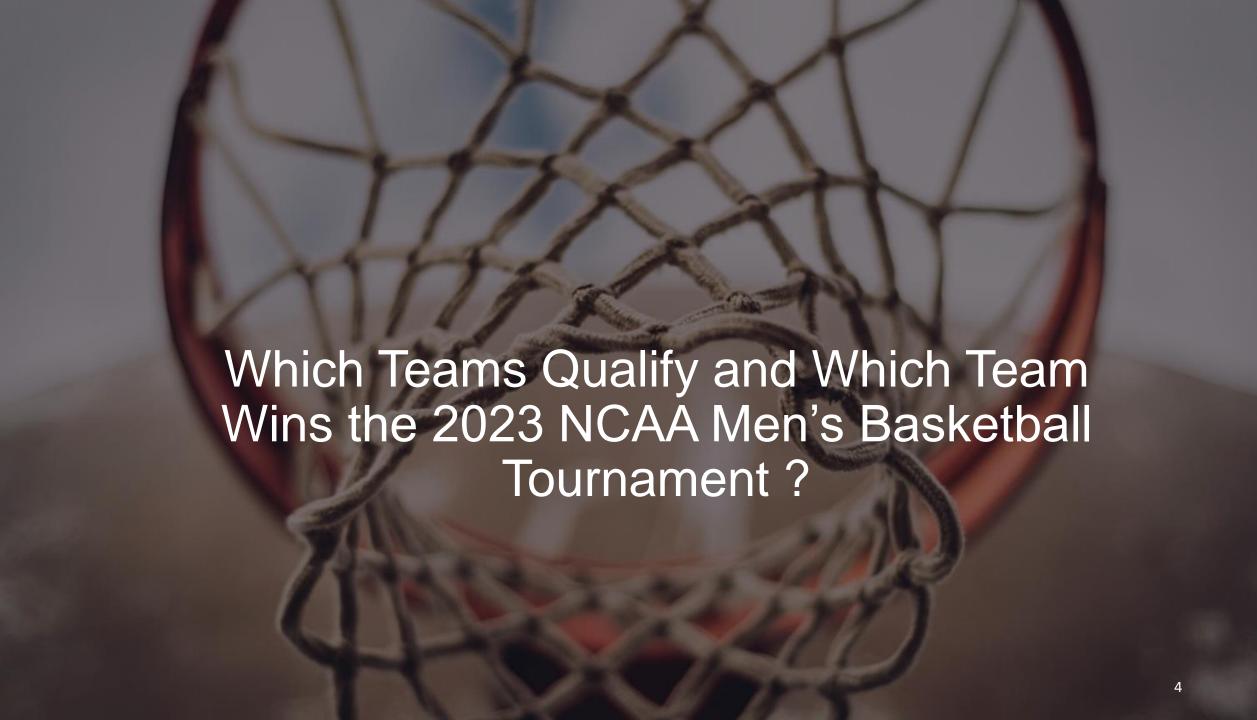
What common characteristics emerge when teams are grouped together in terms of their performance?

K-Means Clustering

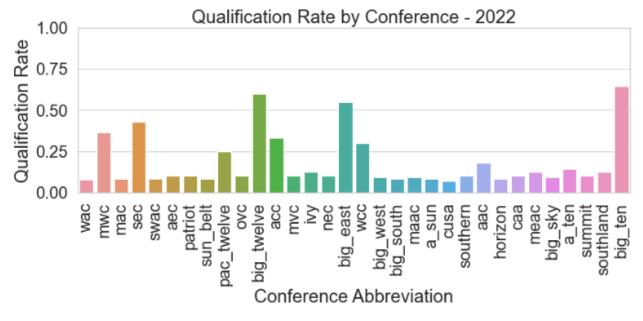
Agglomerative Hierarchical Clustering

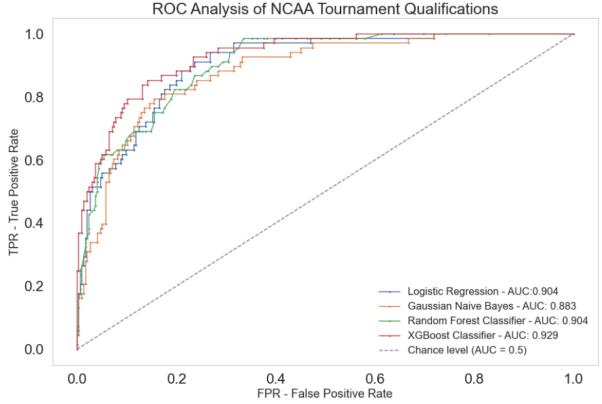
Gaussian Mixture Model

Density-Based Clustering Algorithm (DBSCAN)



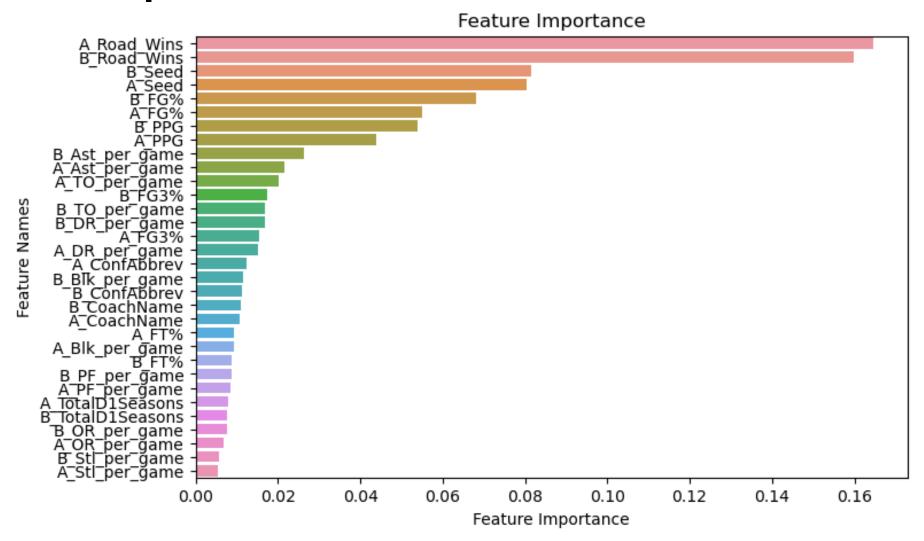
XGBoost Classifier Qualification Results



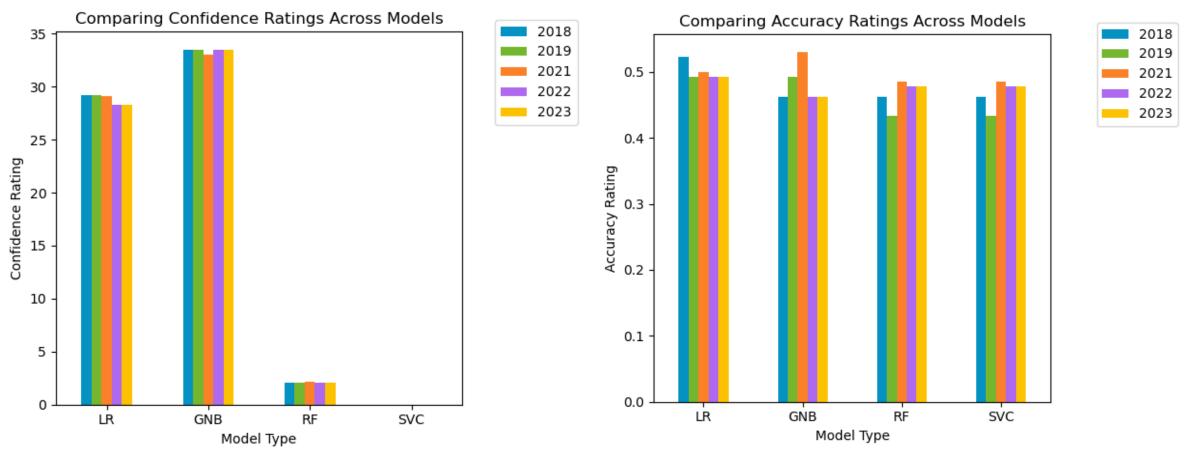




Most Important Features for the Model



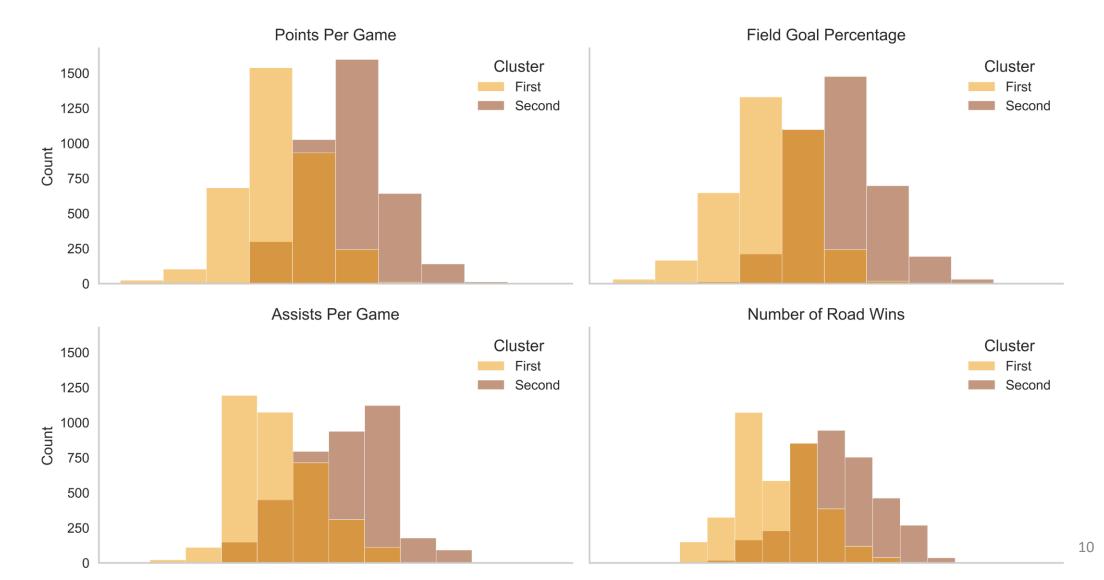
Did Tournament Outcomes Get Harder to Predict?

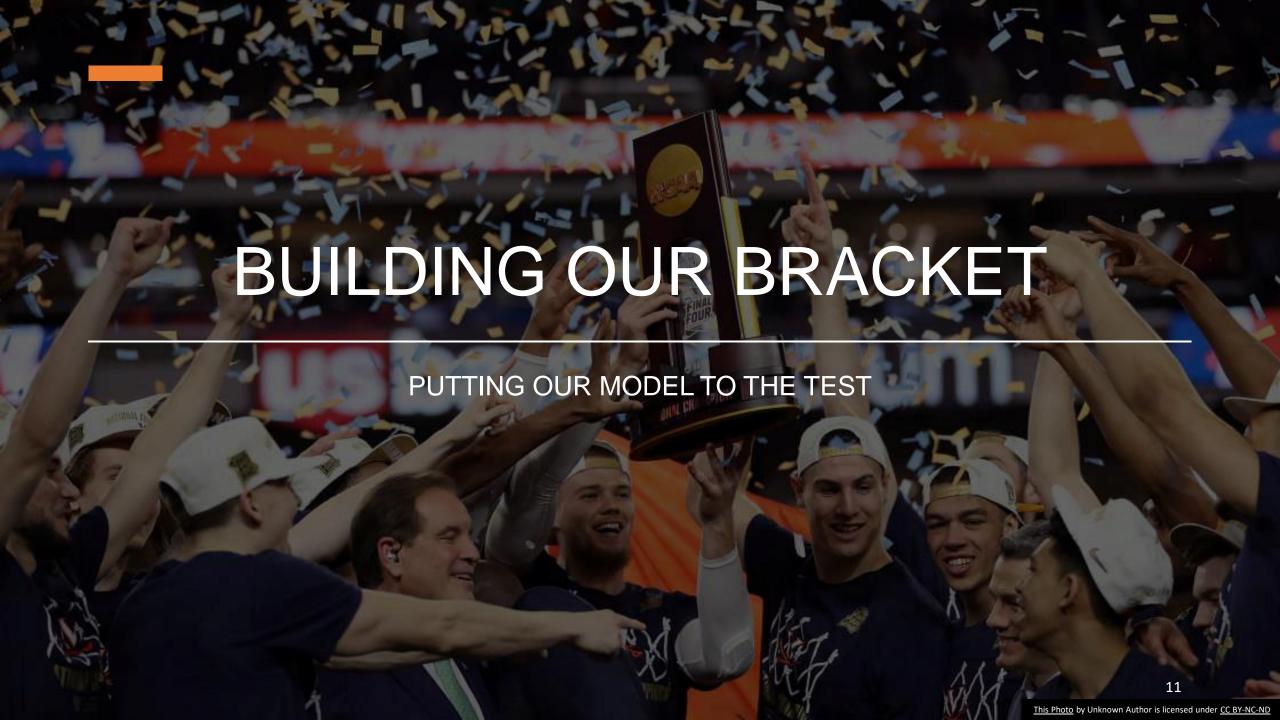


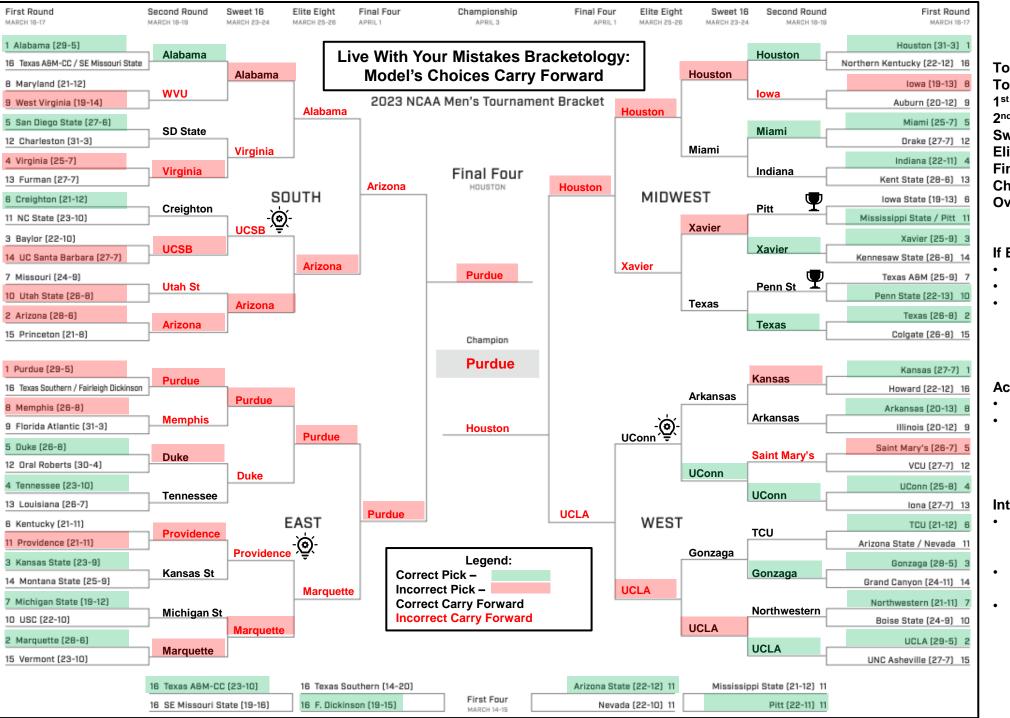
Common Characteristics in Terms of Game Performance When Teams are Grouped Together



K-Means Clustering Results







Game Day Statistics

Total Points Scored – 42
Total Games Correctly Picked – 31

1st Round Accuracy – 69%
2nd Round Accuracy – 50%
Sweet 16 Accuracy – 13%
Elite 8 Accuracy – 0%
Final 4 Accuracy – 0%
Championship Accuracy – 0%
Overall Accuracy – 49%

If Entered in Bracket Competition:

- Average Performance
- Placed 10th in IM Bracket Challenge
- Placed 3rd in Family League



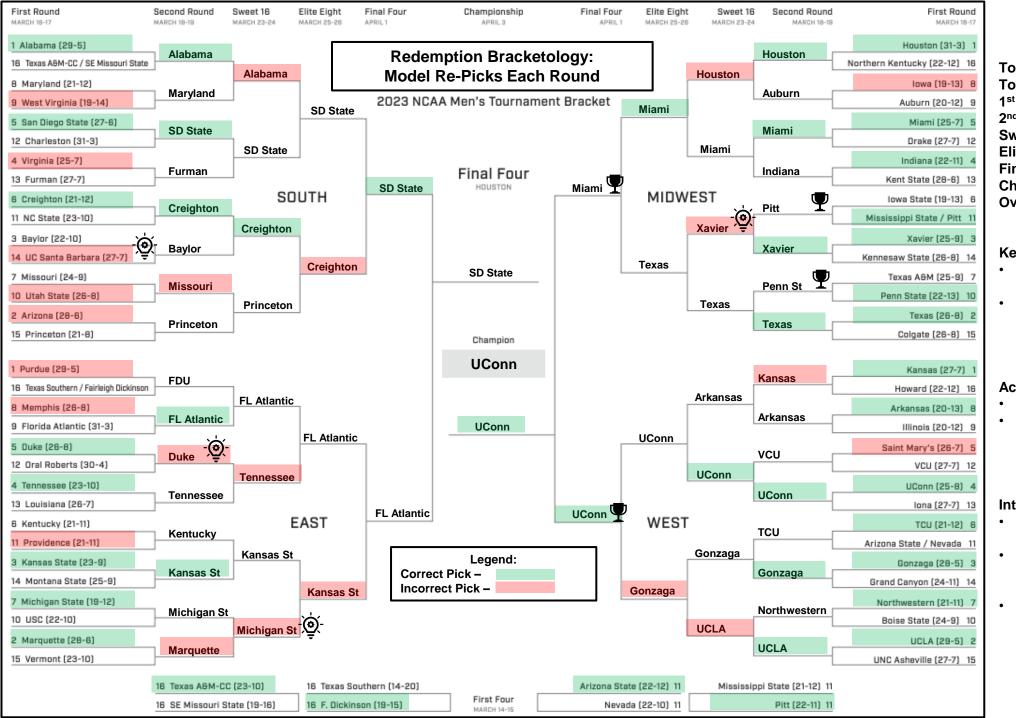
Accurately Picked Upsets:

- Pitt (11) upsets Iowa State (6)
- Penn State (10) upsets Texas A&M (7)



Interesting Insights:

- Model had a strange affinity for UCSB (14), which it predicted to go to the Sweet 16
 - Similarly, it predicted that Providence (11) would advance to the Sweet 16
 - It did see something in UConn (4) who went on to win the national championship!
 - Only team not seeded in top 2 selected to advance to the Elite 8



Game Day Statistics

Total Points Scored – 126 (Not Realistic)
Total Games Correctly Picked – 40

1st Round Accuracy – 69%
2nd Round Accuracy – 75%
Sweet 16 Accuracy – 25%
Elite 8 Accuracy – 25%
Final 4 Accuracy – 100%
Championship Accuracy – 100%
Overall Accuracy – 63%

Key Takeaways:

- Does much better when picking individual game outcomes
 - Does seem to favor seeding, but not exclusively



Accurately Picked Upsets (Additional):

- Miami (5) upsets Texas (2)
- UConn (4) upsets Gonzaga (3)



Interesting Insights:

- Did the model really like UCSB or did it identify Baylor as a very weak #3 seed?
- Picked Michigan St to lose to #2
 Marquette, but, in subsequent game,
 picked them to beat #3 Kansas St
- Tended to pick close upsets for historically strong programs
 - #5 Duke to beat #4 Tennessee
 - #3 Xavier to beat #2 Texas

Conclusions, Discussions and Future Work



Conclusions and Future Work







Classification

- Best classifier: XGBOOST CLASSIFIER
- Worst Classifier: Gaussian Naïve Bayes
- Best Features: Road Wins, PPG, Assists per game and FG%
- Conference Affiliation seems to have some importance in qualification process
- Future Work: Taking individual player data into consideration, especially in the NIL era.

Prediction

- Best classifier: LOGISTIC REGRESSION
 - Worst Classifier: SVC
- Best Features: Road Wins, Seed, FG%, PPG
- We did see a small decrease in overall confidence rating across years with the LR
- The SVC did not generate a probability that deviated from 0.5, would explore alternative models with better probability forecasting

Clustering

- Best clustering method: KMEANS
- Worst clustering method: GMM & DBSCAN
 - 2 clusters found: lower-performing and higher-performing teams
- Common features for best teams : Road wins, PPG, and DR per game
- Future Work: Try new dimensionality reduction method (Manifold Learning)

