2025 Wharton High School Data Science Competition

Machine Learning Approaches to Basketball Analysis

By The Quakers

Shawn Qiu, Ryan Zhang, Tingray Chung, Jude Al-Mufti

Advisor: Jerry Zupan
Horace Greeley High School, Chappaqua, NY

March 2025

Introduction & Background

Task

Use women's basketball data to rank the top 16 teams of each region and predict the winning probabilities of theoretical matchups

Significance

Women's basketball is booming, with 2024 finals viewership at 18.9M vs. 14.8M for the men's game

Traditional Methods and Limitations

Win-Loss Records: Oversimplification of team performance

Logistic Regression: Always assumes linear relationships

Power Rankings: Often subjective and prone to bias

Our Goal

Develop a combined ranking system and a multi-variable predictive model to address these limitations

Methods: Preprocessing

01 DATA CLEANING

- Removed non-D1 team games
- Imputed NAs with 0 or mean
- Applied log transformations and min-max scaling to normalize data

02 AGGREGATION

- Aggregated statistics by team (e.g., average scores)
- Merged regional data

03 ELO SYSTEM

- Iterated through each game to calculate and finalize every team’s Elo rating

04 FEATURE ENGINEERING

- Merged home/away team data into single rows
- Computed difference-based features for the model

Methods: Ranking with K-Means Clustering

01

Unsupervised Learning

Model learns the clusters based on team data without labels

02

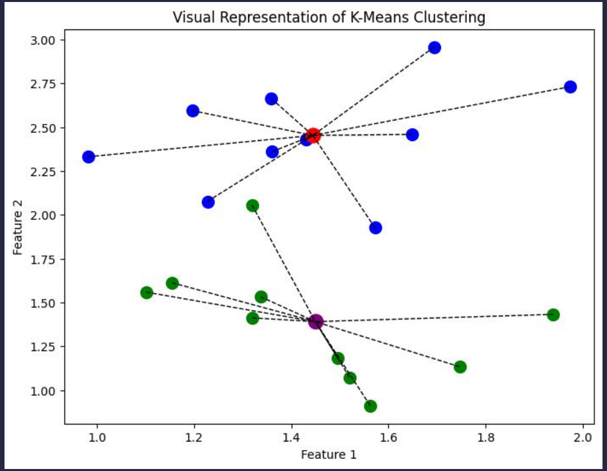
K Number Determination

Used Elbow Method and Silhouette Analysis to select the best K number

03

Euclidean Distance Ranking

Ranked teams by calculating the Euclidean distance of each team to a centroid



Centroid scores calculated by summing the mean feature values for each cluster

Methods: XGBoost for Winning Probabilities

01 **Gbtree Gradient Booster**

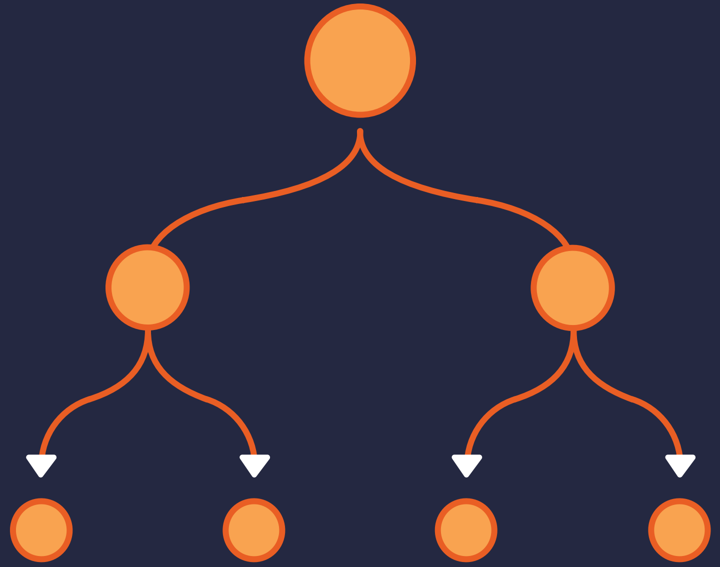
Binary:Logistic objective to calculate winning probabilities

02 **Hyperparameter Tuning**

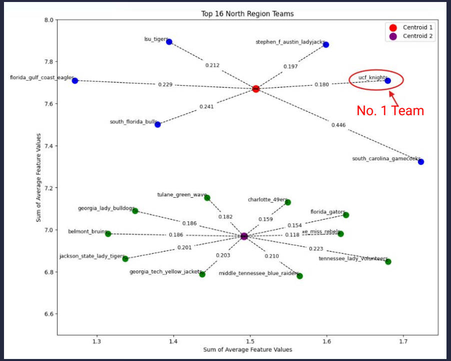
Used grid search cross-validation to identify optimal parameters

03 **Early Stopping Rounds**

Controlled overfitting with adjustment of boosted round quantity



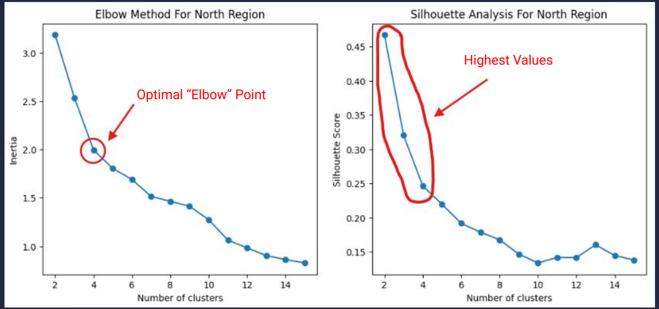
Results: Top 16 K-means Rankings



No. 1 Team

- Assigned teams to the nearest centroid
- Ranked centroids by their calculated score
- Ranked teams by proximity to the centroid in each cluster

Note: only the top 2 clusters are shown



Optimal "Elbow" Point

Highest Values

-Choose a K at the "Elbow"

-Choose a K with a high silhouette score

-The best K for North Region is 4

North Region

Best K Number for the Region	4

Centroid Scores: 7.71, 6.98, 5.99, 5.42

South Region

Best K Number for the Region	3

Centroid Scores: 7.41, 6.53, 5.62

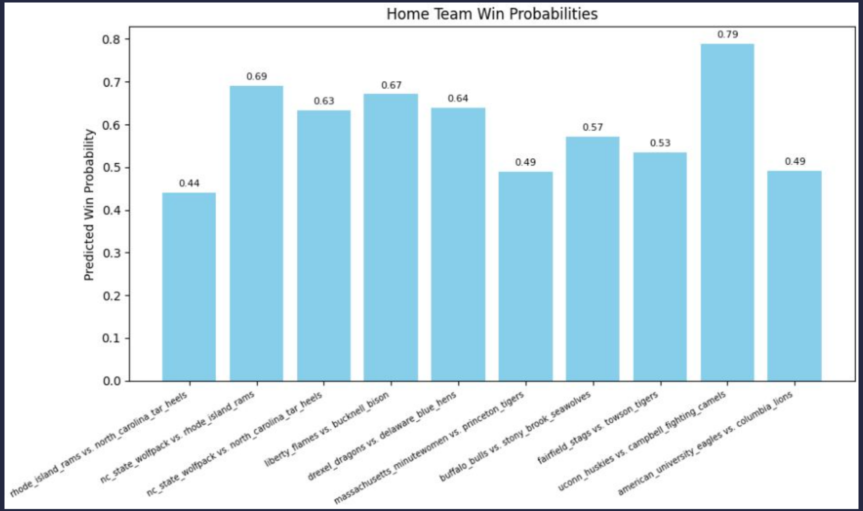
West Region

Best K Number for the Region	4

Centroid Scores: 7.76, 6.61, 5.87, 5.20

Results: XGBoost Winning Probabilities

Model Predictions



Model Performance

Eval AUC Score	0.88
Accuracy Score	0.80

Model Parameters

Learning Rate	0.1
Max Depth	5
Boosted Rounds	100

Conclusion

Utilized K-means clustering

- **With holistic scores calculated from multi-dimensional performance metrics**
- **Achieved rankings for the top 16 teams in each region**

Identified XGBoost as a robust model

- **To capture non-linear relationships among multiple variables**
- **Predicted team winning probabilities**

Our methods offer accurate, adaptable, and holistic evaluations of basketball team performance

Limitations

- We did not include the impacts of certain variables (e.g., attendance, time zone difference, previous game distance) in our analysis
- We were unable to effectively normalize some variables due to their irregularity (e.g., technical foul)

References

- Brown, Bryce, "Predictive Analytics for College Basketball: Using Logistic Regression for Determining the Outcome of a Game" (2019). Honors Theses and Capstones. 475. https://scholars.unh.edu/honors/475
- Matt Gifford, Tuncay Bayrak, A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression, Decision Analytics Journal, Volume 8, 2023, 100296, ISSN 2772-6622, https://doi.org/10.1016/j.dajour.2023.100296.
- Ziv, G., Lidor, R., & Arnon, M. (2010). Predicting team rankings in basketball: The questionable use of on-court performance statistics. International Journal of Performance Analysis in Sport, 10(2), 103–114. https://doi.org/10.1080/24748668.2010.11868506
- Slide template courtesy of https://slidesgo.com/

Acknowledgements – Wharton High School Data Science Competition Team