Explainable Basketball Athlete Profiling

Anant Kedia AU2220040
Maanit Shah AU2220043
Samarth Bankar AU2220001
Viraj Bhatia AU2220038
Ayush Chaurasia AU2220032

***Abstract*—This project aims to propose a comprehensive and explainable framework for profiling basketball athletes using machine learning. The study leverages game data from the North-east Conference (NEC) and Metro Atlantic Athletic Conference (MAAC) from 2020 to 2024, focusing on clustering and predictive modelling. K-means and Fuzzy C-means (FCM) clustering were applied to identify player profiles based on key performance metrics. FCM provided soft membership scores, aligning with the fluid nature of modern basketball positions. Additionally, Random Forest was used to predict game outcomes from the athlete statistics.**

***Index Terms*—Machine learning, Sports analytics, Clustering algorithms, Predictive modelling**

I. Introduction

In the ever-evolving landscape of sports analytics in basketball, machine learning-based approaches have emerged as promising tools for enhancing athlete evaluation and team strategy development [1]. Studies in the field have demonstrated the efficacy of machine learning algorithms in predicting sports outcomes and assessing player performance.

Multiple recent studies have demonstrated the aformentioned. For instance, Wang et al. [2] utilise various different ML approaches to provide a comprehensive evaluation of their performance in predicting game outcomes. However, the inputs in their approach fail to capture real-world factors like team dynamics. Similarly, Chen et al. [3] use a hybrid model that combines various approaches including K-Nearest Neighbours (KNN), Extreme Gradient Boosting (XGBoost) and Stochastic Gradient Boosting (SGB). While this approach improves accuracy, the model's complexity increases, adversely affecting interpretability.

Several approaches have been explored to tackle the problem of explainability. Ouyang et al. [4] integrate Shapley Additive Explanations (SHAP) with XGBoost as an effective approach to improve explainability. However, this integration does increase the computational complexity and fails to generalise performance indicators. Islam et al. [5] propose iXGB, an approach to enhance XGBoost's interpretability by approximating decision rules from its internal structure and generating counterfactuals. Nevertheless, the approximation of decision rules may not capture all the complexities of the model.

All of the approaches discussed also fail at capturing some real-world factors. Hu et al. [6] quantify player interactions at synergies to enhance the XGBoost prediction model. However, quantifying synergies for a larger set of players is difficult. Also, the method is extremely sensitive to outliers.

This project focuses on tackling some of the limitations discussed above. The main objective of this project is to propose a comprehensive and explainable framework for profiling basketball athletes.

II. Methodology

We used a dataset consisting of the Northeast Conference (NEC) and Metro Atlantic Athletic Conference (MAAC) matches from 25/11/20 to 25/03/24. These conferences are affiliated with the National Collegiate Athletic Association (NAAC) Division-1 [7] [8]. Each match is represented by the two teams playing it, and the teams' respective information such as the athletes participating in the match and standard box statistics representing their performance. Furthermore, the teams' total score in the match and whether they won or lost is present in the data. Finally, an aggregate score of each athlete's performance, called Hollinger's Game Score, is present, calculated as

$$\begin{aligned} Game\ Score = PTS + (0.4 \cdot FGM) - (0.7 \cdot FGA) \\ -\big(0.4 \cdot (FTA - FTM)\big) + (0.7 \cdot OREB) \\ + (0.3 \cdot DREB) + STL + (0.7 \cdot AST) \\ + (0.7 \cdot BLK) - (0.4 \cdot PF) - TO, \end{aligned} \tag{1}$$

where the abbreviations are the standard basketball box scores.

A. Clustering Approaches

We aim to classify athletes into discrete disjoint groups via unsupervised clustering. This will allow us to infer about possible categorisations of the athletes based on their individual performances. We perform the following clusterings on the box statistics of individual athletes averaged over all of their appearances in matches.

1) K-Means Clustering: We use k-means clustering as a preliminary unsupervised clustering algorithm. Applying it to the dataset as is, with 16 total parameters, results in the following plot for the Within Cluster Sum of Squares (WCSS) estimate.

Following the "elbow method" (Fig. 1), we find the $k = 11$ value to be optimal. Furthermore, we perform dimensionality reduction via Principle Component Analysis (PCA) to consider only 2 components/parameters so as to reduce the effects of the Curse of Dimensionality.

2) Fuzzy C-means: We used Fuzzy C-Means (FCM) clustering to obtain soft position memberships, as player positions are not rigidly defined, and athletes usually take on multiple roles.

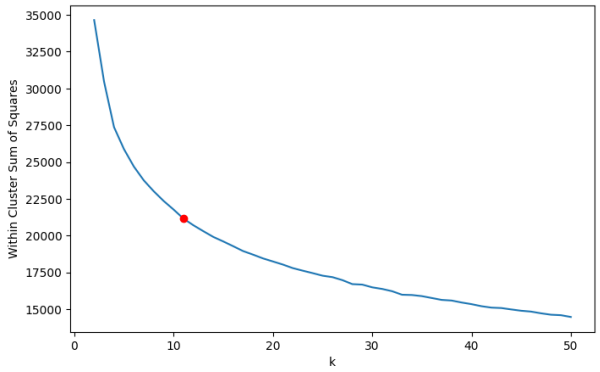

Fig. 1. WCSS vs k for the elbow method

We first extracted seven key performance metrics from the dataset: Field Goal Percentage (FG%), Points Scored (PTS), Three-Point Percentage (3P%), Assists (AST), Rebounds (REB), Steals (STL), and Blocks (BLK). Two of these statistics, FG% and 3P%, were calculated using the formulas:

$$\text{FG\%} = \frac{\text{FGM}}{\text{FGA}}, \quad \text{3P\%} = \frac{\text{3PM}}{\text{3PA}} \tag{2}$$

Given a membership matrix U and weight matrix W from FCM clustering, where u_{ij} represents the degree to which player i belongs to cluster j, we computed position-specific memberships using:

$$P = UW \tag{3}$$

This transformation ensures that each player has a soft membership score for all five basketball positions, reflecting their suitability for multiple roles.

B. Game Outcome Prediction Using Random Forest

One of the primary objectives of the project is game outcome prediction. A game outcome predictor will help coaches come up with the optimal line-up for the team and help identify potential strengths and weaknesses in both competing teams.

We employed the Random Forest algorithm to predict game outcomes based on individual athletes' statistics. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes or mean prediction of the individual tree.

The dataset was partitioned using an 80-20 split for training and testing. The model was configured with 100 decision trees to balance complexity and generalisability. We applied a random forest prediction model to the dataset on individual athletes' statistics (averaged over all their appearances) and trained the model over their number of wins.

III. Results

1) K-Means Clustering: We received poor results on analysing the k-means clustering model via unsupervised clustering metrics. Specifically, the model had a Silhouette score of 0.14 and Davies-Bouldin index of 1.87. Similarly, the model applied to the dataset post-PCA processing had poor results.

A. Fuzzy C-Means

We then applied Fuzzy C-Means clustering to group players into five clusters, corresponding to the five traditional basketball positions: Point Guard (PG), Shooting Guard (SG), Small Forward (SF), Power Forward (PF), and Center (C). However, from the five clusters generated, four had meaningful centroids and an additional cluster emerged, consisting of players with minimal impact. To map the clusters to meaningful basketball positions, we constructed a weight matrix based on domain knowledge and cluster centroids. By analysing the statistical tendencies of each cluster's centroid, we observed that some clusters leaned more toward specific positions. For example, clusters with high assists and moderate scoring were associated with PGs, while clusters with strong rebounding and blocking tendencies were linked to PFs and Cs. This knowledge guided the formation of the following position weight matrix:

$$W = \begin{bmatrix} 0.1 & 0.5 & 0.4 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2 & 0.5 & 0.3 \\ 0.2 & 0.3 & 0.4 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.6 & 0.1 & 0.2 & 0.1 & 0.0 \end{bmatrix} \tag{4}$$

Using W, the position-specific membership scores were calculated for each player.

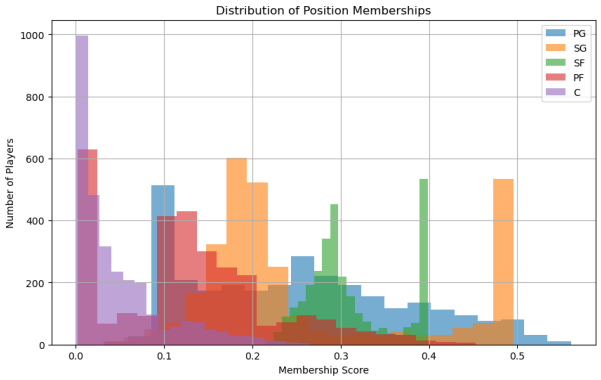

Fig. 2. Membership Histogram for the 5 Positions

B. Random Forest

The model had an R^2 value of 0.45, which is undesirable. This is most definitely due to the fact that match wins are a result of team efforts, and the statistics of individual athletes are not enough to predict them. Feature importance analysis revealed that defensive statistics, particularly blocks (BLK) and steals (STL), were the most significant predictors of win probability.

IV. Discussion

The application of clustering algorithms, specifically K-means, to the dataset yielded challenges due to its high dimensionality. This complexity likely hindered the algorithm's

ability to form distinct clusters, suggesting that dimensionality reduction techniques or alternative clustering methods may be necessary to achieve more meaningful groupings.

The integration of Fuzzy C-means clustering provided soft membership scores for traditional basketball positions. This approach aligns with the fluid nature of modern basketball, where players often exhibit skills across multiple positions. The resulting profiles offer a more nuanced understanding of player capabilities, which could assist recruiters in identifying talent that fits specific team dynamics.

Predictive modelling using Random Forests to forecast match outcomes resulted in an R^2 value of 0.45, indicating limited predictive power. This outcome underscores the inherent complexity of basketball, where game results are influenced by numerous factors beyond individual player statistics, such as team dynamics, coaching strategies, and in-game variables. Future models might improve accuracy by incorporating these additional dimensions.

The planned exploration of XGBoost and the use of explainability tools like SHAP and LIME represent promising directions for future research. XGBoost's ability to handle high-dimensional data and model complex interactions could enhance predictive performance. Simultaneously, SHAP and LIME can provide insights into feature importance, offering a clearer understanding of the factors driving model predictions.

V. Conclusions

Even though the results we got were quite undesirable, the implementations made helped create a solid base for the ultimate aim of this project, i.e. to create an explainable and comprehensive framework for profiling basketball athletes. While the focus of most studies in the area remains the accuracy of the model, we believe that finding a balance between accuracy and explainability is essential, as it helps coaches, scouts and recruiters who might not have a background in computer science make better strategic decisions based on data.

References

[1] R. P. Bunker and F. Thabtah, "A machine learning framework for sport result prediction," *Applied Computing and Informatics*, vol. 15, no. 1, pp. 27–33, 2019.

[2] J. Wang, "Predictive analysis of nba game outcomes through machine learning," 01 2024, pp. 46–55.

[3] W.-J. Chen, M.-J. Jhou, T.-S. Lee, and C.-J. Lu, "Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association," *Entropy*, vol. 23, no. 4, 2021.

[4] Y. Ouyang, X. Li, W. Zhou, W. Hong, W. Zheng, F. Qi, and L. Peng, "Integration of machine learning xgboost and shap models for nba game outcome prediction and quantitative analysis methodology," *PLOS ONE*, vol. 19, no. 7, pp. 1–25, 07 2024. [Online]. Available: https://doi.org/10.1371/journal.pone.0307478

[5] M. R. Islam, M. U. Ahmed, and S. Begum, "iXGB: Improving the interpretability of XGBoost using decision rules and counterfactuals," in *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024)*, vol. 3, 2024, pp. 1345–1353. [Online]. Available: https://www.scitepress.org/Papers/2024/124740/124740.pdf

[6] H. Hu, G. Dimitrov, D. Menn, and S. Wu, "NBA player performance prediction based on XGBoost and synergies," in *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022. [Online]. Available: https://courses.cs.washington.edu/courses/cse547/23wi/old_projects/23wi/NBA_Performance.pdf

[7] "The Northeast conference." [Online]. Available: https://northeastconference.org/sports/2024/8/22/gen_aboutnec_2425.aspx

[8] "The MAAC." [Online]. Available: https://maacsports.com/sports/2017/6/20/GEN_0620175706.aspx