Athlete Profiling: NCAA

CSE632: Machine Learning Theory and Practise

Dhaivat Patel	Dhyey Patel	Sloka Thakkar	Anusha Jain	Tirthraj Raval
AU2240022	AU2240054	AU2240103	AU2240092	AU2240079

***Abstract*—Athlete profiling is a crucial aspect of modern sports analytics, aiding recruiters in identifying players who best fit team strategies. This study introduces a data-driven framework for classifying NCAA women's basketball players based on their offensive and defensive ratings. Using K-Means clustering, players are grouped into offensive, defensive, and neutral roles, providing a more nuanced understanding of their contributions. A logistic regression model is then applied to classify new data points. The results demonstrate high classification accuracy, though class imbalance affects minority class predictions. This approach offers recruiters deeper insights into player strengths and weaknesses, enhancing talent identification and team composition.**

***Index Terms*—athete profiling, team profiling, logistic regression, k-means clustering,**

I. Introduction

Profiling athletes as well as evaluating their performance is fundamental to modern sports analytics because it helps recruiters and coaches choose players who best fit the team tactics. Traditional measures, such as game scores, tend to miss an athlete's more delicate contributions. This investigation proposes a comprehensive structure within which Offensive Rating (ORTG) and Defensive Rating (DRTG) are combined to evaluate a player's skill and a team's productivity. Utilizing data pertaining to NCAA women's basketball for the last four years, we classify athletes using these ratings for the benefit of recruiters. This method provides multi-faceted analysis of athletes, surpassing basic scoring systems to accentuate deeper capabilities and deficiencies.

Our framework is based on two factors - offensive and defensive ratings, rather than a player's holistic game score. A combination of these components is more informative for estimating individual player contribution and for measuring team contribution to the players' performance. In addition, our method enables us to assess the team's performance. Though both factors matter, team productivity is mostly determined by offense, as poor offensive performance reduces productivity.

The major outputs of this undertaking comprise: - Defining on a definite scheme a team's strengths and weaknesses in both scoring and non scoring.

II. Methodology

A. Dataset Discussion

The data set contains 23 features of basket ball players (Team Display Name, Athlete, Date, Opponet, PTS, MIN, FGM, FGA, PM, PA, FTM, FTA, OREB, DREB, REB, AST, BLK, STL, TO, PF, Team Score, Win, Game Score). There are 28388 rows.

Identify applicable funding agency here. If none, delete this.

B. Player's Offensive and Defensive Rating

Basketball players can be generally classified into three categories: offensive players (attackers), defensive players (defenders), and neutral players. In order to classify players into these categories it is required to measure their offensive and defensivecontri- butions quantitatively. This is done by computing offensive Identify applicable funding agency here. If none, delete this. rating (ORTG) and defensive rating (DRTG), which yield a numeric measure of a player's offensive and defensiveplay efficiency, respectively.

$$ORtg = 100 \times \left(\frac{\text{PProd}}{\text{TotPoss}} \right) \tag{1}$$

Where, PProd represents the total number of points a player contributes, including field goals, assists, and free throws.
And TotPoss accounts for the total number of possessions a player is involved in, including field goal attempts, turnovers, and free throw trips.

$$DRtg = D_1 + D_2 \tag{2}$$

D1 represents the opponent's scoring efficiency per 100 possessions.
D2 adjusts for defensive contributions by factoring in stops, opponent efficiency, and free throw impact.

C. K-Means Clustering

After ORTG and DRTG scores are calculated, players must be categorized accordingly. As the dataset does not have predefined labels for offensive, defensive, or neutral players, an unsuper- vised learning method is used.
Initially, the players with average game play minutes less than 10 minutes were eliminated leaving behind 2285 rows. The scores are projected in two-dimensional space, and clustering is done through K-Means. The algorithm creates three clusters, categorizing players as Offensive, Defensive, or Neutral on the basis of their performance statistics.

D. Logistic Regression

After training the model and classifying players into three groups, logistic regression is applied to classify new data points. The model learns weight parameters during training, which are subsequently used to classify players in the testing dataset.
This ensures that each player is assigned to the most suitable category based on their performance metrics

III. Results

After computing the offensive and defensive scores for each player, we applied clustering techniques to categorize them into three distinct groups: offensive, defensive, and neutral players.

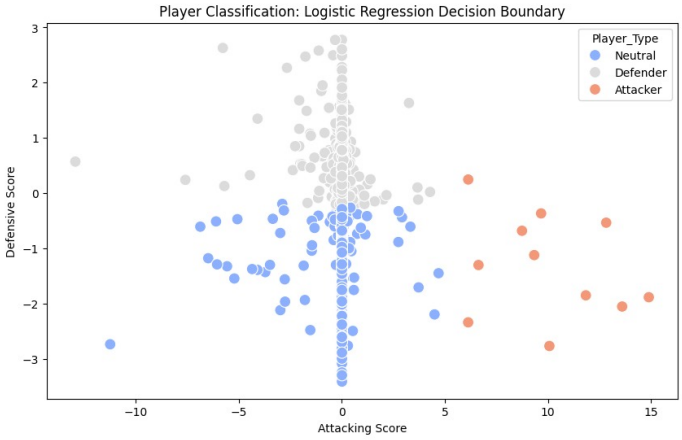

Fig. 1. Graph after K means Clustering.

Following the clustering process, we employed Logistic Regression for further analysis.

Class	Precision	Recall	F1-score	Support
0	0.00	0.00	0.00	1
1	1.00	1.00	1.00	289
2	0.99	1.00	1.00	167
Accuracy		1.00		457
Macro Avg	0.66	0.67	0.67	457
Weighted Avg	1.00	1.00	1.00	457

TABLE I
Classification Report

IV. Discussion

In the clustering results, we observe that the majority of players fall into the neutral category. This could be because the neutral category includes players who primarily play in central positions and exhibit both offensive and defensive skills. Additionally, it may also include players who are primarily offensive or defensive but did not perform up to the expected level in their respective matches.

1) Class 0: The single sample in this class was completely misclassified as Class 2. This misclassification is primarily due to a severe class imbalance.

- **Precision (0.00):** No instances of Class 0 were correctly predicted.
- **Recall (0.00):** The only sample belonging to Class 0 was misclassified.
- **F1-score (0.00):** Since both precision and recall are zero, the F1-score is also zero.

2) Class 1: All 289 samples were correctly classified as Class 1.

- **Precision, Recall, and F1-score (1.00, 1.00, 1.00):** The classification for this class was perfect.

3) Class 2: The majority of samples (167) were correctly classified as Class 2.

- **Precision (0.99):** A near-perfect score, likely due to one sample being misclassified as Class 1.
- **Recall and F1-score (1.00, 1.00):** The model effectively identified all instances of Class 2.

Overall, **Classes 1 and 2 were classified with high accuracy**, whereas **Class 0 suffered from complete misclassification** due to its severe underrepresentation in the dataset. Addressing the class imbalance through techniques such as resampling or weighted loss functions could improve the model's performance for minority classes.

Conclusion

This study introduced a data-driven approach to athlete profiling in NCAA women's basketball, utilizing offensive and defensive ratings to categorize players. By applying K-Means clustering, we effectively grouped players into offensive, defensive, and neutral roles. The subsequent logistic regression model demonstrated high accuracy in classifying new data points, though class imbalance affected minority class predictions. Our findings highlight the importance of multi-dimensional performance analysis, providing recruiters with deeper insights into player strengths and weaknesses. Future work could explore advanced classification techniques and address dataset imbalances to enhance predictive accuracy.

References

[1] Calculating individual offensive and defensive ratings — Basketball-Reference.com. (n.d.). Basketball-Reference.com. https://www.basketball-reference.com/about/ratings.html
[2] Author links open overlay panelVangelis Sarlis et al., "Sports analytics - evaluation of basketball players and Team Performance," Information Systems, https://www.sciencedirect.com/science/article/pii/S0306437920300557 (accessed Mar. 17, 2025).
[3] D. Oliver, Basketball on Paper: Rules and Tools for Performance Analysis. S.l.: Potomac Books, 2020.
[4] M. Á. Pérez-Toledano, F. J. Rodriguez, J. García-Rubio, and S. J. Ibañez, "Players' selection for basketball teams, through performance index rating, using Multiobjective Evolutionary Algorithms," PLOS ONE, https://doi.org/10.1371/journal.pone.0221258 (accessed Mar. 17, 2025).