NCAA Basketball Analytics: ML-Powered Player Insights and Optimal Lineups

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 the contribution of the individual player and for measuring the contribution of the team to the performance of the players. 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.

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

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.

B. Player's Offensive and Defensive Rating

Generally, Basketball players can be 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 defensive contributions 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 defensive play 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. Player classification using 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

E. Win-Loss Prediction

1. Linear Regression with Game Score

To predict match outcomes, player-level data including offensive rating (ORtg), defensive rating (DRtg), and game score were added to the match dataset. Players with less than 5 minutes of playtime were excluded. For each team, these stats were averaged to compute team-level offensive, defensive, and game scores. The same was done for the opponent. These features were used to train a Linear Regression model to predict match results (win or loss).

2. Linear Regression without considering Game Score

To analyze match outcomes without using game score, player-level offensive (ORtg) and defensive (DRtg) ratings were added to the dataset. Players with less than 5 minutes of playtime were removed. Team-level offensive and defensive scores were calculated by averaging player stats, and the same was done for opponent teams. A Linear Regression model was trained on these features to predict match results. The model achieved.

3. Linear Regression using Game Score Only

In this approach, a Linear Regression model was trained using only team-level Game Scores to predict match outcomes. Game Score was calculated by summing the scores of individual players (who played at least 5 minutes) and averaging them per team. Both the team's and the opponent's average Game Scores were used as features.

4. Comparative Modeling for Win prediction using Ratings and Game Score

In this experiment, three machine learning models—Logistic Regression, Random Forest, and XGBoost—were evaluated using a combined feature set of team-level average ratings: Offensive Rating (ORtg), Defensive Rating (DRtg), and Game Score for both teams. Only players with at least 5 minutes of playtime were included, and all metrics were normalized by the number of players per team. All models performed well, with XGBoost achieving the highest accuracy. These results indicate that combining ORtg, DRtg, and Game Score provides a strong feature representation for predicting match outcomes.

5. Comparative modeling for win prediction using Game Score

In this study, game score—were was used to predict match outcomes. Only players with at least 5 minutes of playtime were included. For each team, a Team_Game_Score was calculated by aggregating player Game Scores and normalizing by the number of players. The opponent's average Game Score, Team_Game_Score_Opp, was also included. The final feature set consisted of:

The binary target variable, Win, represented match outcomes (1 for win, 0 for loss). Three classification models were trained: Logistic Regression, Random Forest, and XGBoost, with an 80-20 train-test split.

6. Match Outcome Prediction using ORtg and DRtg

This task aimed to predict match outcomes based on team performance, using player-level ratings and match statistics. Player ratings (Offensive Rating—ORtg, Defensive Rating—DRtg) were calculated from the parameters given in the dataset. The equation was used from the book, Oliver Dean: Basketball on Paper. A function was implemented to merge player ratings with match data using athlete names. Only players with at least 5 minutes of playtime were retained. The Win column was converted into binary format (1 for win, 0 for loss). Opponent team metrics were also added to capture relative performance. Infinite values were replaced with NaN, and rows with missing values were dropped.
Three models—Logistic Regression, Random Forest, and XGBoost—were trained on an 80-20 train-test split.

F. Optimal Lineup Prediction using Reinforcement Learning

To predict the optimal lineup of the basketball for maximum probability of winning the match, reinforcement learning was used. Offensive Rating(ORtg) and Defensive Rating(DRtg). Logistic Regression model(as above) was used to predict win or loss of the match based on the given opponent players and predict players using the reinforcement learning model. For reinforcement model, Q-learning was used, which iteratively gets optimal 5-player lineups against a fixed opponent team. The predicted lineup is than used to predict what is the win probability using the Logistic regression model.

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. For team analysis, after model learning, the accuracies and loss scores were compared to evaluate the overall results of model. For team lineup, we used the team analysis and it gives the best 5 plyaer suitable that maximises the winning changes, given which 5 players ofthe opponent team will be plyaing.

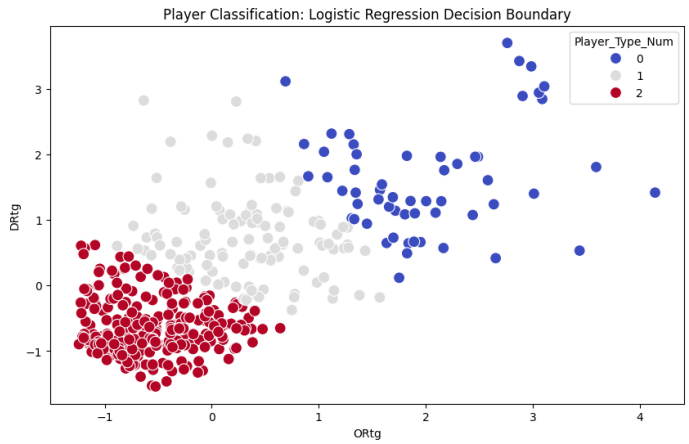

Fig. 1. Graph depicting the results of K-means clustering, where 0 represents Attackers, 1 represents Defenders, and 2 represents Neutral players.

TABLE I
CLASSIFICATION REPORT

Class	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	8
1	1.00	0.96	0.98	28
2	0.98	1.00	0.99	53
Accuracy		0.99		89
Macro Avg	0.99	0.99	0.99	89
Weighted Avg	0.99	0.99	0.99	89

TABLE II
MODEL ACCURACY WITH ORTG AND DRTG

Model	**Accuracy**
Logistic Regression	0.7500
Random Forest	0.6995
XGBoost	0.7197

TABLE III
MODEL ACCURACY WITH ONLY ORTG, DRTG AND GAME SCORE

Model	**Accuracy**
Logistic Regression	0.9242
Random Forest	0.9141
XGBoost	0.9268

TABLE IV
MODEL ACCURACY WITH ONLY GAME SCORE

Model	**Accuracy**
Logistic Regression	0.9272
Random Forest	0.8988
XGBoost	0.8988

IV. DISCUSSION

Discussion of categorizing offensive and defensive player-Player Analysis

In the clustering results, we observe that Class 2 (Neutral players) comprises the majority of the dataset. This may be attributed to the fact that neutral players often operate in central roles, contributing both offensively and defensively. Additionally, some players who typically serve as attackers or defenders but underperformed in specific matches might have been grouped into this neutral cluster due to their average performance in both metrics.

1) Class 0: This class, representing Attackers, had 8 samples, all of which were correctly classified.

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

2) Class 1: This class represents Defenders, with 28 samples. One sample was misclassified.

- **Precision (1.00):** All samples predicted as Class 1 were indeed Class 1.
- **Recall (0.96):** One instance was incorrectly classified, slightly lowering the recall.
- **F1-score (0.98):** High overall performance with a small trade-off in recall.

3) Class 2: This class represents Neutral players, the largest group with 53 samples.

- **Precision (0.98):** A few misclassifications led to a small dip in precision.
- **Recall and F1-score (1.00, 0.99):** The model identified all actual Class 2 instances correctly.

Overall, Classes 0 and 1 were classified with excellent accuracy, while Class 2 showed minor misclassification due to its larger size and overlap with other roles. Future work could benefit from exploring feature engineering or hierarchical classification to better distinguish nuanced player roles.

Discussion of win prediction-Team Analysis

The findings draw attention to a crucial finding regarding the Game Score metric. Points scored and other game variables are used to compute the game score, a statistic that is determined after a game. Because of the large variation in points across games, it is extremely changeable even though it offers an overall performance metric. This fluctuation may mask significant trends that affect the results of matches. A more consistent and dependable method of analyzing team performance is provided by the use of player ratings, such as Offensive Rating (ORtg) and Defensive Rating (DRtg). These ratings are more accurate predictors of match results since they are based on the contributions of individual players and are not as affected by chance variations.

Random Forest, XGBoost, and Logistic Regression were all assessed in terms of model performance. With the highest accuracy (92.72%), Logistic Regression appears to be a good fit for the goal of binary categorization of match outcomes. When the correlations between features and the target variable

are linear, Logistic Regression is a strong candidate for this kind of prediction because to its ease of use and excellent interpret-ability. Random Forest and XGBoost, on the other hand, performed marginally worse even though they also achieved good accuracy. These models capture more intricate, non-linear correlations and interactions between features since they are ensemble approaches. Since the data in this instance does not show extremely complicated interactions that would necessitate the complexity of Random Forest or XGBoost, Logistic Regression is the best model to use.
Furthermore, the simplified model requires less computing power and provides a clear interpretation of feature importance, both of which are useful in sports analytics to comprehend the main factors influencing match results.

Discussion on Optimal Team Lineup

This study demonstrates that combining logistic regression with Q-learning offers an effective method for basketball lineup optimization. The logistic regression model, trained on team-level metrics like ORtg and DRtg, provided reliable win probability estimates that served as rewards in a reinforcement learning environment. The Q-learning agent successfully identified lineups that maximized predicted win probabilities, enabling data-driven selection based on both offensive and defensive contributions. This approach moves beyond static strategies by allowing adaptive exploration of optimal player combinations. However, the linear nature of logistic regression may not capture complex player interactions, and the model's accuracy depends on the quality of input ratings. Future work could explore non-linear models and incorporate real-time data or contextual features to enhance performance.

V. 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. Using the offensive and defensive ratings it become and using linear regression on it, the coaches can know their possibility of winning and train the players accordingly. Moreover, the optimal lineup model using reinforcement learning helps the coach to send the players which can have high probability of winning the match against the opponent. Thus, 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] Basketball-Reference.com, "Calculating individual offensive and defensive ratings," [Online]. Available: https://www.basketball-reference.com/about/ratings.html [Accessed: Apr. 13, 2025].

[2] D. Oliver, *Basketball on Paper: Rules and Tools for Performance Analysis*, Potomac Books, 2020.

[3] "How to analyze basketball player performance and predict MVP Awards," PlotsAlot, https://plotsalot.slashml.com/blogs/basketball-sports-analyticsmvp-prediction-results (accessed Apr. 13, 2025).

[4] J. Poropudas and T. Halme, "Dean Oliver's four factors revisited," arXiv.org, https://doi.org/10.48550/arXiv.2305.13032 (accessed Apr. 13, 2025).

[5] M. Á. Pérez-Toledano, F. J. Rodríguez, J. García-Rubio, and S. J. Ibáñez, "Players' selection for basketball teams, through performance index rating, using multiobjective evolutionary algorithms," *PLOS ONE*, vol. 14, no. 8, 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0221258 [Accessed: Mar. 17, 2025].

[6] R. Taylor, "NBA Player Efficiency Rating vs. Winning Percentage," *Taylor University Sport Management Student Projects*, 2015. [Online]. Available: https://pillars.taylor.edu/cgi/viewcontent.cgi?article=1001context=sport-management-student-projects [Accessed: Apr. 13, 2025].

[7] squared2020, "Game score: Focus on scoring," Squared Statistics: Understanding Basketball Analytics, https://squared2020.com/2017/09/20/game-score-focus-on-scoring/ (accessed Apr. 13, 2025).

[8] T. Neiman and Y. Loewenstein, "Reinforcement learning in professional basketball players," Nature News, https://www.nature.com/articles/ncomms1580 (accessed Apr. 13, 2025).

[9] V. Sarlis, D. Amaxilatis, E. Amaxilatis, and A. Tsihrintzis, "Sports analytics – Evaluation of basketball players and team performance," *Information Systems*, vol. 95, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437920300557 [Accessed: Mar. 17, 2025].