Athlete performance in collegiate basketball: Predicting match line-up

Jash Parikh, AU2140108, Nevil Jobanputra, AU2140209, Paridhi Jain, AU2120226, Saahil Doshi, AU2140106

***Abstract*—This paper presents a streamlined approach to enhance athlete performance in collegiate basketball through predictive analytics and visualization techniques. Leveraging a diverse dataset encompassing sleep patterns, training details, emotional states, game scores, and more for Division I basketball players, our study focuses on predicting match line-ups and aiding decision-making.**

Methodologically, we employ data augmentation, time-series analysis for RSImod prediction, and eXplainable AI for result interpretation. Additionally, we develop an algorithm to forecast match line-ups based on key feature scores. Through advanced Machine Learning techniques, particularly RandomForestRegressor models, we achieve high accuracy in RSImod prediction.

Furthermore, we emphasize the development of a visual dashboard using R Shiny, enabling real-time visualization and comparison of individual performance metrics against team averages. This user-friendly interface facilitates coaches in identifying areas for improvement and optimizing training strategies.

By addressing the outlined problem statement, our research offers a concise framework for leveraging data-driven insights to elevate athlete performance and team competitiveness in collegiate basketball. Through the integration of predictive analytics and interactive visualization tools, we aim to revolutionize sports management practices, providing coaches with actionable insights for optimal performance outcomes.

***Index Terms*—Basketball, RSImod, Random Forest Regressor.**

I. Introduction

In the dynamic world of professional sports, predictive analysis has emerged as an essential and impactful means for teams to stay ahead of their opponents.[1] The use of analytics allows teams to get information about their players' strengths and weaknesses as well as team performance. This allows them to make more informed decisions on how they can maximize their capacity and come up with the best strategies. Analytics data can also provide teams with information on game situations which can be used to figure out the opponent's move and consequently adjusting their own strategies accordingly. Analytics can enable the team to discover what needs to be strengthened in the areas of player formation and team cohesion. Moreover, it serves as a tool for the analysis of trends and evaluation of different approaches and methods. In short, analytics has changes the way teams view performance and strategy, allowing them to gain a tactical advantage over their rivals.

Our work stands does not rely on game-play statistics and outcomes of past matches, instead focusing on the physiological and cognitive aspects of athletes. By analyzing sleep patterns, recovery strategies, and cognitive states, we aim to provide a comprehensive understanding of how they can be utilized to give the best match line-up. This approach not only offers a fresh perspective on match line-up but also provides valuable insights for coaches and teams to optimize training and game strategies.

II. Methodology

Data

In this project, we use the data obtained from the season 2 and season 3 performance of Division 1 Basketball players. The Dataset includes sleep patterns, training details, cardiac rhythm patterns, emotional-mental state information, game scores, weekly readiness scores, and jump-data (RSImod) of the atheletes.

Handling Missing Values

In our analysis, we encountered missing data in both season 2 and season 3 Polar dataset. To address this issue and ensure the reliability of our analysis, we applied the Multivariate Imputation by Chained Equation (MICE) method.

MICE is a robust technique that allows us to predict missing values by considering the relationships among variables in the dataset. This approach is particularly effective even when dealing with datasets containing a high proportion of missing values.

By applying MICE to the Polar dataset, we were able to accurately fill in the missing data points, thereby improving the quality of our analysis. This method reduced bias and enhanced the reliability of our results, leading to more reliable insights and conclusions.

Multicollinearity analysis

Multicollinearity analysis highlighted the presence of linear dependencies among features, potentially introducing bias into the ML model's predictions. A simplistic approach of dropping features could compromise the predictions. Thus, in addressing this issue, the utilization of factor analysis (FA) remains pending. FA aims to amalgamate features with similar variances into a condensed, lossless, and alternative representation. By modeling features as a function of latent variables, FA facilitates their combination into smaller groups termed factors. Subsequently, ML techniques are applied to these factors to generate predictions at player, team, and conference levels. In line with the broader domain of ML applications, such as sports science, ensuring fairness, accountability, and transparency in the decisions made by ML models is paramount. One method to elucidate ML decisions is through feature importance analysis, which assigns scores to factors based on their impact on the game score. This ranking system enhances interpretability, offering insights into

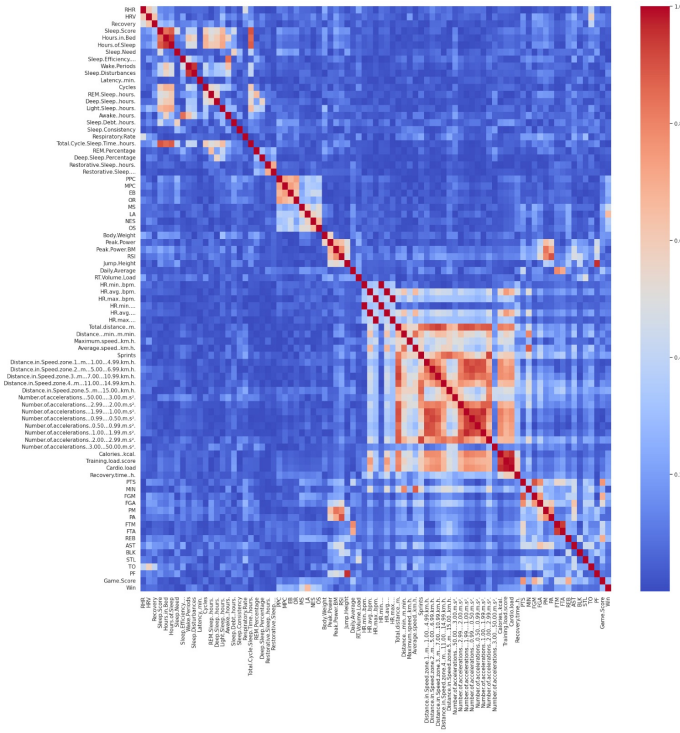

Fig. 1. Correlation Matrix

the model's predictions.

Determining Feature Importance

To ensure consistency and usability of the data, we identified and removed the irrelevant colums, such as athlete names, dates and test types. Next, the features, representing attributes like body weight, peak power, and jump height, are separated from the target variable, RSImod, which measures players' readiness or performance level.

With the obtained data we further trained a machine learning model using the selected features and the RSImod target variable. In this case, a RandomForestRegressor model is utilized for its effectiveness in handling regression tasks. During training, the model learns the underlying patterns and relationships between the features and RSImod. Random Forest constructs multiple decision trees and evaluates the importance of features based on their contribution to reducing impurity in the trees.These importances indicate the relative contribution of each feature to predicting RSImod, providing insights into the factors influencing players' readiness or performance in Division I basketball.

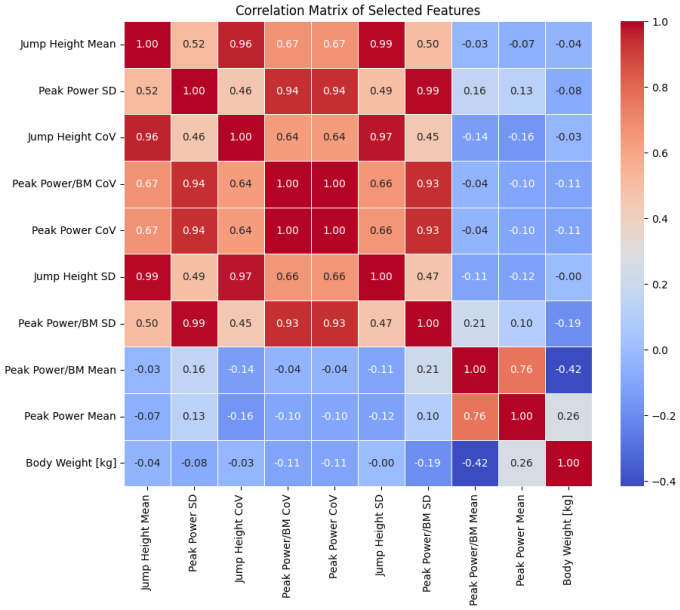

Fig. 2. Feature correlation Matrix

Training the model

The process involves a number of steps in training a predictive model for RSImod that can be applied to a Division I basketball player. At first, the features which are relevant to the dataset were selected and they include jump height, peak power, and body weight, as these features are believed to influence the RSImod. The data once prepared were divided into training and testing sets consisting 80% data trained on the model and the rest 20% data for testing of the model.

We chose RandomForest Regressor model for training because of this model's matching for regression tasks. The model is taught by learning the original features and their RSImod values using the training set, so it can catch the specific patterns and relationships of data. The next step after training is the evaluation of the model performance using the testing set. It is the RSImod predictions generated from the testing features that provides the mean squared error (MSE) to quantify the deviation between the actual and the predicted RSImod values.

III. Discussions

After determining feature importance and training the model, the next step would be evaluating the model's predictive performance using data from both season 2 and season 3 datasets. Specifically, the trained model is utilized to predict RSImod values for the season 3 dataset, treating it as the test dataset whereas data from season 2 is the training dataset. To integrate time series analysis into this process, the chronological order of the data is maintained, ensuring that temporal dependencies are captured appropriately. The model's predictions for RSImod in season 3 are then compared against the actual RSImod values from the same dataset. This comparison allows for the assessment of the model's accuracy in capturing temporal patterns and trends in RSImod values across different seasons. As a simple measure of accuracy, the mean squared error (MSE) can be calculated to quantify the average squared difference between the predicted and actual RSImod values.

IV. Conclusion

Analysing player performance is capable of producing a revolution in sports management and training systems. It can be used to analyze data from previous performances and therefore help coaches make an informed choice on team strategy and player selection. Machine Learning can be used effectively in monitoring the on-field progress of players in

real-time, supplying the coaches with a very useful information on the players' strengths and weaknesses.

V. References

[1] The Rise of Predictive Analytics in Professional Sports. Daily Press, 2023. website: https://www.dailypress.net/sponsored-content/2023/04/the-rise-of-predictive-analytics-in-professional-sports/.

[2] Taber, C.B., Sharma, S., Raval, M.S. et al. A holistic approach to performance prediction in collegiate athletics: player, team, and conference perspectives. Sci Rep 14, 1162 (2024). https://doi.org/10.1038/s41598-024-51658-8

[3] S. U. Sharma, S. Divakaran, T. Kaya and M. Raval, "A Hybrid Approach for Interpretable Game Performance Prediction in Basketball," 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022, pp. 01-08, doi: 10.1109/IJCNN55064.2022.9892583.

[4] Claudino, J.G., Capanema, D.d., de Souza, T.V., Serrão, J.C., Machado Pereira, A.C., Nassis, G.P., et al. (2019). Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review. Sports Medicine - Open, 5, Article number: 28. https://doi.org/10.1186/s40798-019-0202-3.

[5] Wang, Yuanchen Liu, Weibo Liu, Xiaohui. (2022). Explainable AI techniques with application to NBA gameplay prediction. Neurocomputing. 483. 59-71. 10.1016/j.neucom.2022.01.098.