American Journal of Applied Mathematics and Statistics, 2024, Vol. 12, No. 4, 75-79
Available online at http://pubs.sciepub.com/ajams/12/4/1
Published by Science and Education Publishing
DOI:10.12691/ajams-12-4-1

Analysis and Prediction of CBA Player Position Data Characteristics Based on Machine Learning

Yuanzheng Yu*

Hamden Hall Country Day School, Hamden, USA
*Corresponding author: yuanzhengyu.jack@gmail.com

Received October 05, 2024; Revised November 07, 2024; Accepted November 14, 2024

Abstract This study uses player performance statistics from the Chinese Basketball Association for six seasons, from 2017 to 2022, to evaluate the statistical characteristics of guards, forwards, and centers. 20 key performance indicators including points per game, rebounds, assists, shooting percentages, etc. are employed to provide empirical evidence to identify the singular traits that have come to be associated with each position. The study uses eight different machine learning models -- Decision Tree, Linear Discriminant Analysis, Multinomial Logistic Regression, Naive Bayes, Neural Network, Random Forest, Support Vector Machine, and XGBoost -- for position prediction of players from their performance data. From the results, it can be learned that guards are much more adept at scoring, assists, steals, and three-point shooting; forwards are better rebounders and three-point shooters; centers are proficient in rebounding, blocking, and field goal percentage. Among all the considered predictive models, Random Forest and XGBoost have the best test accuracies, while some models are clearly overfitted. This study suggests that using an ensemble machine-learning approach on performance data in the CBA context works particularly well when predicting player's position. The study contributes to a better understanding of positional attributes in professional basketball and provides methodological references for future research in the field of sports analytics.

Keywords: *CBA (chinese basketball association), player position analysis, machine learning classification, sports analytics, performance metrics*

Cite This Article: Yuanzheng Yu, "Analysis and Prediction of CBA Player Position Data Characteristics Based on Machine Learning." *American Journal of Applied Mathematics and Statistics*, vol. 12, no. 4 (2024): 75-79. doi: 10.12691/ajams-12-4-1.

1. Introduction

The CBA is the highest professional basketball league in China, presenting talented players from various parts of the country and abroad. To gain proper insights for effective team strategy, player development, and spectator engagement, players' distinctive features need to be understood as classified by their playing position-guard, forward, or center-on the court. Traditionally, position descriptions have been based more on folklore and tradition rather than on scientific study. The advances in statistical methodologies and machine learning approaches provide an opportunity for such positional attributes to be studied more rigorously.

Despite the considerable body of research focused on analytics within the NBA and Euroleague [1], there exists a notable deficiency in studies pertaining to the Chinese Basketball Association (CBA). The objective of this investigation is to remedy this shortcoming by employing machine learning methodologies on CBA datasets, drawing upon approaches utilized in earlier basketball analytics scholarship. Nevertheless, despite the popularity of CBA, comprehensive studies that statistically test the differences between positions are scant. That would be critical for providing empirical evidence on the characteristics associated with guards, forwards, and centers. Moreover, it remains unclear which algorithms in machine learning are best in predicting players based on the performance statistics of the CBA.

Player position analysis in basketball has been one of the important research areas for determining different responsibilities and contributions of guards, forwards, and centers within a team framework. Traditionally, studies have depended on subjective assessments of expert evaluations and game observations to define the responsibilities typically associated with each position [2]. More recent efforts have utilized advanced statistical metrics and performance analytics to delineate the positional role in greater detail. Research employing box score metrics, including Player Efficiency Rating (PER) and Win Shares, has provided numerical insights into the performance of athletes across various positions [3]. Moreover, spatial analytics and tracking data have permitted scholars to examine player movements and interactions on the court, thereby enhancing the quantitative differentiation among positions [4]. Strategies of basketball development, especially the increasing usage of the so-called "small-ball" lineup, have dictated a reevaluation of traditional positional roles. This evolution underlines the need for flexible and adaptive player

appraisal, as evidenced by recent studies which explore hybrid positions and the convergence of positional responsibilities across positions during games such as Fu et al., 2021 [5]. Although substantial research has been conducted regarding player position analysis for the NBA and Euroleague, studies related to the CBA are relatively few. The aim of this work is to fill this gap by applying machine learning techniques to CBA data.

Machine learning (ML) has fundamentally transformed the field of sports analytics, providing advanced methodologies for predictive modeling, pattern identification, and assistance in decision-making processes. Specifically in basketball, ML methodologies have been utilized to predict player performance, refine team strategies, and improve scouting operations (Bunke & Susnjak, 2022). A range of machine learning models, such as Decision Trees, Support Vector Machines (SVM), Random Forests, and Neural Networks, have been employed to categorize player positions, forecast game results, and pinpoint significant performance metrics [6]. For instance, Random Forest and Gradient Boosting Machines have been leveraged to anticipate player injuries by examining past performance and biometric information [7]. Techniques of unsupervised learning, including clustering and dimensionality reduction, have been employed in the segmentation of players according to performance attributes and the discovery of concealed patterns within team dynamics [4]. Deep learning frameworks, specifically Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been utilized to scrutinize video recordings, facilitating the automated identification of player actions and tactical formations [5]. The analysis by Albert et al. in 2022 used some machine learning algorithms, including regularized linear regression and random forest, to predict rankings within NBA teams and salaries for players. Their findings demonstrated how the approach is valuable for basketball analytics, specifically in the context of NBA analysis [8]. However, how this could relate to other leagues, such as the CBA, remains a topic of further research.

Research specific to the CBA has been relatively limited compared to studies that focus on the NBA. Wang et al. conducted an exploratory analysis of player efficiency in the CBA, highlighting the differences in scoring patterns and defensive metrics compared to NBA players. Results from their study showed that CBA players are more proficient in three-point shooting and free-throw shooting, which might be due to differences in coaching styles or the way each coaching staff approaches the game. Wang et al. focused on the impact of foreign players on the performance dynamics of CBA teams. Utilizing regression models, the research demonstrated that teams with a higher proportion of foreign players tend to have improved offensive efficiency, particularly in scoring and assist metrics [9]. Tan et al. explored the relationship between player longevity and performance sustainability in the CBA [10]. Their longitudinal study analyzed career trajectories of CBA players over a decade, identifying key factors that contribute to sustained high performance, such as versatility in playing multiple positions and adaptability to evolving team strategies. In a recent study focused specifically on the CBA, Kang and Xu, 2020 utilized weighted linear regression to predict CBA team rankings, achieving an accuracy of 61.4% [11]. Their work provides a foundation for applying machine learning techniques to CBA data and demonstrates the potential for predictive analytics in this league. However, their study also highlights the need for more comprehensive analyses that incorporate a wider range of statistical features and machine learning models.

This paper is focused on analyzing the statistical characteristics of guards, forwards, and centers in the CBA over six seasons from 2017 to 2022, offering a quantitative framework for understanding the roles of these positions. To achieve this, various machine learning models were employed to predict players' positions based on their performance statistics, with an emphasis on determining which algorithms are most effective for this classification task. By identifying the models with the best predictive accuracies, this research aims to propose an optimal method for player position classification using CBA data, while simultaneously contributing to the advancement of sports analytics.The results of this research provide data-driven insights into the positional attributes of CBA players, offering significant implications for coaches and team managers regarding player assessment, training priorities, and strategic planning. By assessing the performance of various machine learning algorithms in predicting player positions, this study highlights the best methodologies for applying data analytics in sports, thereby enhancing predictive modeling techniques that can be applied in similar contexts. Building on previous studies, this work extends the use of machine learning models for player position prediction in the CBA, incorporating a broader range of statistical features. In doing so, the study contributes to the expanding literature on basketball analytics while providing insights that are specifically tailored to the CBA context.

2. Methodology

2.1. Data Description

Data for this analysis are sourced from https://basketball.realgm.com and covers a six-season span of the Chinese Basketball Association: from 2017 to 2022. The dataset gives an overview of players in 20 CBA teams with 2066 player-season records in total.

Data Cleaning Process:

1. Position Classifications: These are the various positions which the players play in the court: guards, forwards, and centers. At positioning, hybrid positions include point guards who predominantly perform duties of typical shooting guards.

2. Redundancy Removal: We detected and removed 52 duplicate entries to assure the integrity of the data.

3. Handling Cases of Repeated Names: For 37 cases where a player's name appeared several times during a single season, each was treated separately to account for possible team roster changes during that season or multiple data entry.

The performance indicators used in this study are listed in Table 1.

Table 1. Performance indicators used in this study

Abbreviation	Description	Abbreviation	Description
GP	Games Played	FTA	Free Throws Attempted
MPG	Minutes per Game	FT%	Free Throw Percentage
PPG	Points per Game	ORB	Offensive Rebounds
FGM	Field Goals Made	DRB	Defensive Rebounds
FGA	Field Goals Attempted	RPG	Rebounds Per Game
FG%	Field Goal Percentage	APG	Assists Per Game
3PM	Three-Point Field Goals Made	SPG	Steals Per Game
3PA	Three-Point Field Goals Attempted	BPG	Blocks Per Game
3P%	Three-Point Field Goal Percentage	TOV	Turnovers
FTM	Free Throws Made	PF	Personal Fouls

2.2. Classification Models

We used eight machine learning models to classify the player's position:

1. **Decision Tree**: It is used to identify key features that influence position classification.
2. **Linear Discriminant Analysis**: This serves as a basic linear classification model.
3. **Multinomial Logistic Regression**: It is used to predict the probability of each position class.
4. **Naive Bayes**: It is the probabilistic baseline classifier in this paper.
5. **Neural Network**: It contains a multilayer perceptron with three hidden layers each with 100 neurons, using the ReLU for activation and lbfgs (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) optimizer.
6. **Random Forest**: The number of estimators is taken to be 100.
7. **Support Vector Machine**: The radial basis function RBF kernel is used.
8. **XGBoost**: It is the most popular boosting classifier.

All models were developed in Python 3.9 using the scikit-learn library 1.0.2 and the XGBoost library 2.0.3. The train-test-split is 70% vs 30%.

3. Results

3.1. Descriptive Statistics

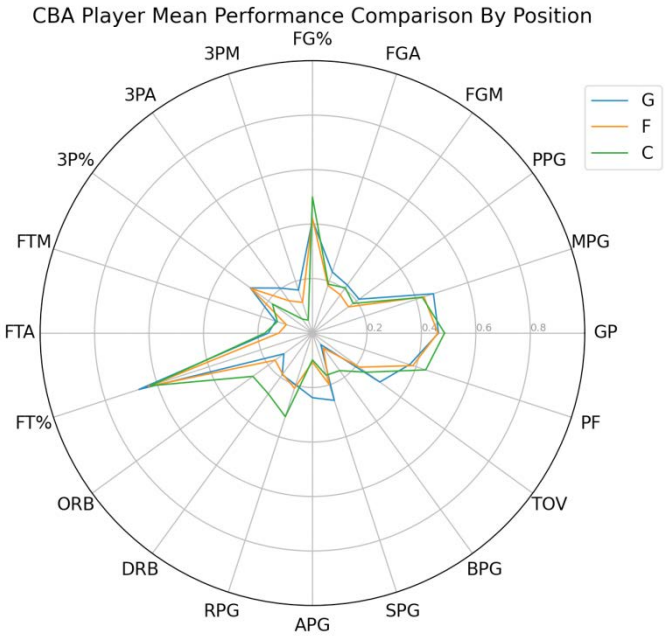

Figure 1. Illustrates the radar chart of three position-averaged performance indicators of the CBA players. Table 2 shows the summary statistics (mean and std) of the performance indicators for the three positions: guard, forward and center

Table 2. Summary Statistics of Performance Indicators of Guard, Forward and Center

	Guard		Forward		Center	
	mean	std	mean	std	mean	std
GP	27.85	15.86	27.72	15.67	29.14	14.99
MPG	20.13	10.64	18.56	9.27	18.29	9.25
PPG	9.34	8.76	7.22	5.90	8.16	6.68
FGM	3.33	3.05	2.66	2.18	3.14	2.57
FGA	7.52	6.31	5.86	4.18	6.02	4.54
FG%	0.41	0.13	0.42	0.13	0.50	0.14
3PM	1.03	1.10	0.73	0.73	0.31	0.53
3PA	3.00	2.90	2.15	1.85	0.94	1.40
3P%	0.28	0.16	0.28	0.17	0.18	0.20
FTM	1.65	1.95	1.17	1.31	1.57	1.53

	Guard		Forward		Center	
FTA	2.12	2.35	1.62	1.68	2.29	2.13
FT%	0.67	0.26	0.63	0.26	0.63	0.22
ORB	0.75	0.74	0.98	0.85	1.57	1.13
DRB	2.27	1.94	2.24	1.86	3.31	2.63
RPG	3.02	2.53	3.22	2.56	4.87	3.65
APG	2.61	2.37	1.17	1.10	1.08	1.16
SPG	0.86	0.66	0.65	0.49	0.54	0.43
BPG	0.20	0.37	0.26	0.36	0.65	0.59
TOV	1.44	0.98	1.00	0.65	1.15	0.82
PF	1.88	0.86	1.94	0.89	2.19	0.87

3.2. Model Evaluation

We performed the classification of position based on the performance indicators with eight different machine learning algorithms. All the models were trained and tested excluding the feature 'Year', considering only the performance features of the players. Table 3 presents the performance metrics for each model, including training accuracy, test accuracy, precision, recall, and F1-score.

Table 3. Model Performance Metrics %

Model	Train Acc	Test Acc	Precision	Recall	F1-Score
Random Forest	*96.68*	*69.19*	69.42	*69.19*	*68.96*
XGBoost	96.54	68.87	68.88	68.87	68.79
SVM	69.64	67.90	*70.09*	67.90	67.71
Neural Network	96.54	66.29	65.70	66.29	65.87
LDA	63.21	64.35	66.18	64.35	63.72
Logistic Reg	64.38	64.19	64.89	64.19	63.70
Decision Tree	*96.68*	57.42	56.95	57.42	57.11
Naive Bayes	55.19	53.39	58.00	53.39	51.53

Random Forest reached the best result of the highest test accuracy at 69.19%, while XGBoost reached 68.87%. In fact, the best results of both have proven that they can really model the data in complicated patterns. Support Vector Machine (SVM) test accuracy is 67.90%, with the max precision 70.09%, which is quite good and hence reflects the appropriateness of this method for rightly estimating player position classes. Neural Network and Linear Discriminant Analysis (LDA) showed moderate test accuracies of 66.29% and 64.35%, respectively. Logistic Regression-LDA also did almost equally well, correctly classifying 64.19% of the test cases. Decision Tree and Naive Bayes models had lower test accuracies of 57.42% and 53.39%, respectively, indicating less effectiveness in this classification task.

A striking thing in all these models is how different the training and test accuracies are. Random Forest, XGBoost, Neural Network, and Decision Tree all had very high training accuracies above 96%, but their test accuracies were considerably lower. That would mean these models are overfitting, most likely learning noise and certain patterns from the training data that generalize very poorly to new, unseen data. SVM has a relatively smaller difference from its training accuracy 69.64% compared to its test accuracy 67.90%, indicating better generalization and less over-fitting. LDA and Logistic Regression also displayed minimal differences between training and test accuracies, reflecting their simpler model structures and tendency to generalize better in certain contexts.

Besides the training accuracy and test accuracy, we also presented the evaluation method of precision, recall, and F1-score. Precision measures the proportion of correctly predicted positive observations to total predicted positives.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall or sensitivity is the ratio of correctly predicted positive observations to all actual positives.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

F1-Score is the weighted average of precision and recall, hence giving a balance between the two.

$$\mathrm{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The Random Forest model demonstrated a precision of 69.42% and a recall rate of 69.19%, culminating in an F1-score of 68.96%. Such equilibrium suggests reliable performance across various assessment criteria. XGBoost also did very well: precision and recall were around 68.88%, combined with an F1-score of 68.79%. SVM performed the classification for the classes marked as positive, with the highest accuracy at 70.09%. Neural Network showed slightly lower precision and recall compared to both Random Forest and XGBoost. This can potentially be due to overfitting, as the high training accuracy is contrasted with the lower test accuracy. Decision Tree and Naive Bayes had the lowest F1-scores, which further confirms their low accuracies.

Performance metrics analysis shows that the ensemble methods, such as Random Forest and XGBoost, outperformed other models in predicting the position of players in the CBA. However, high accuracies during training suggest that both models may be overfitting to the dataset. While the Support Vector Machine (SVM) already presents reduced training accuracy, still comparable test accuracy with high precision suggests that good generalization capabilities exist, meaning an SVM might be a better choice if one wants to have reliable performance on data that has not been seen yet. Less complex models like LDA and Logistic Regression perform moderately; while their performance does not achieve the highest accuracy rates, it does have limited

overfitting, which, in certain contexts, can be useful and thus allows them to keep the same stable accuracy regardless of the addition of the year feature.

4. Discussion and Conclusion

This paper represents the comprehensive work of classification analysis of the player's position in the Chinese Basketball Association from 2017 to 2022, using machine learning algorithms in order to forecast the player position by performance statistics.

Insights drawn from this study have many practical applications. The identified key performance indicators for each position can guide targeted training programs. For instance, the data suggests that improving forwards' three-point shooting could be a valuable focus, given their already competitive accuracy compared to guards. The high three-point shooting percentage of the forwards leads to potential "small ball" lineups without sacrificing any perimeter shooting. Coaches can try to get lineups that maximize this shooting ability by retaining rebounding strength.

While this study offers valuable insights, several limitations should be noted for future research. First, although the six-season span is significant, it may not fully capture long-term trends, and a follow-up study should extend the analysis, potentially covering data from the league's inception to the present. Second, while the 20-feature model used is comprehensive, it does not account for all aspects of player performance. Incorporating more sophisticated metrics such as Player Efficiency Rating, Win Shares, and spatial data like shot locations would enhance the model's predictive power. Addressing these limitations and expanding the approach will lay the groundwork for more advanced analyses of playing styles and performances in the CBA, thereby contributing to the broader field of international basketball analytics.

References

[1] Mandić, R., Jakovljević, S., Erčulj, F., & Štrumbelj, E. (2019). Trends in NBA and Euroleague basketball: Analysis and comparison of statistical data from 2000 to 2017. PloS One, 14(10), e0223524.

[2] Drinkwater, E. J., Pyne, D. B., & McKenna, M. J. (2008). Design and interpretation of anthropometric and fitness testing of basketball players. Sports Medicine, 38(7), 565-578.

[3] Gutiérrez, G. (2024). Performance analysis in basketball players. International Journal of Sports, Exercise and Physical Education, 6(1), 86-87.

[4] Kalman, S., & Bosch, J. (2021). NBA Lineup Analysis on Clustered Player Tendencies: A New Approach to the Positions of Basketball & Modeling Lineup Efficiency. MIT Sloan Sports Analytics Conference.

[5] Fu, X. B., Yue, S. L., & Pan, D. Y. (2021). Tracking and detection of basketball movements using multi-feature data fusion and hybrid YOLO-T2LSTM network. Soft Computing.

[6] Musa, R. M., Majeed, A. P. P. A., Kosni, N. A., & Abdullah, M. R. (2022). Machine Learning in Team Sports: Performance Analysis and Talent Identification in Beach Soccer & Sepak-takraw. Springer, Cham.

[7] Rossi, A., Pappalardo, L., & Cintia, P. (2022). A narrative review for a machine learning application in sports: An example based on injury forecasting in soccer. Sports, 10(5), 5.

[8] Albert, A. A., de Mingo López, L. F., Allbright, K., & Gomez Blas, N. (2022). A Hybrid Machine Learning Model for Predicting USA NBA All-Stars. Electronics, 11(1), 97.

[9] Wang, X., Han, B., Zhang, S., Zhang, L., Lorenzo Calvo, A., & Gomez, M.-Á. (2022). The differences in the performance profiles between native and foreign players in the Chinese Basketball Association. Frontiers in Psychology, 12, Article 788498.

[10] Tan, H., Qin, C., Ran, Z., Wang, K., & Zheng, Z. (2024). Python-based visualization platform implementation of CBA players' regular season 2022-2023 data. Frontiers in Humanities and Social Sciences, 4(7), 104-118.

[11] Kang, N., & Xu, H. (2020). A Study of CBA League Team's Winning Probability Based on 'Grey Prediction'. International Journal of Innovation in Science and Mathematics, 8(3), 124-129.

© The Author(s) 2024. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).