Predicting NBA Shot Success Using Machine Learning

Thesis Paper for I-492 Project

by

Rylen Grundy

Advisor
Dr. Sridhar Ramachandran

TABLE OF CONTENTS

I. Abstract

This study explores the use of machine learning to predict the success of shots taken in the National Basketball Association (NBA) based on game context and player-specific features. The research hypothesized that models trained on a large, multi-season dataset would outperform random guessing and identify key variables that influence shot success. A dataset of over 4 million shot attempts from the 2003–2004 to 2023–2024 NBA seasons was used to train and evaluate two models: logistic regression and gradient boosting. After preprocessing the data, including the use of one-hot and k-fold encoding techniques, both models were assessed for accuracy and feature importance. The gradient boosting model achieved an accuracy of 64%, slightly outperforming the logistic regression model at 63%. The most influential predictors were shot type, shot distance, season, and player name. These findings suggest that both contextual and player-level features play significant roles in determining shot outcomes. While the model's accuracy reflects the inherent variability of live game scenarios, it successfully identifies patterns in player behavior and shooting trends across seasons. This research contributes to the growing field of basketball analytics and provides a foundation for developing more advanced, interpretable shot prediction models in the future.

II. Introduction

Basketball, like many professional sports, has undergone a transformation in recent years, driven by the increasing role of data analytics. Nowhere is this shift more apparent than in the National Basketball Association (NBA), where widespread adoption of tracking technology and statistical modeling has changed how players are evaluated, how strategies are developed by coaches, and how players play the game. Teams no longer rely solely on a coach's intuition or a player's reputation. Decisions about shot selection, play-calling, and player development are now heavily influenced by data. The use of data in basketball has the potential to unlock new heights that the sport has never seen before.

With this potential comes tangible innovation that has already transpired. Analytics in basketball have been used to show shooting trends throughout the history of the league, quantify the difference in shooting skill across generations of players, and determine which statistics from players and teams attribute to winning games. One task has still been unsolved though due to its unpredictable nature — predicting whether a shot will go in or not. Though there are existing studies that have made their fair share of attempts, my project attempts to build upon their work and contribute to the unsolvable. The goal of my project is to develop a machine learning model that predicts the success of NBA shots more accurately than a random guess, based on game context and player-specific features. I also aim to identify which features are most important when attempting to predict shot success. With this model, future analysts will be one step closer to accurately predicting the success of a single shot. The future perfect model would allow players to know what a good shot is, for coaches to be certain in the offensive and defensive strategies they implement, and for front offices to be able to create teams that fit their specific game plan and style of play. The contributions of this project would be a major stepping stone in the field of analytics in basketball.

III. Literature Review

Over the last two decades, the integration of analytics into basketball has fundamentally changed the sport. This shift is especially visible in the NBA, where data-driven decision-making has become central to everything from roster construction to in-game strategies. Siddique (2024) demonstrated how machine learning and statistical modeling can optimize team lineups within constraints such as the salary cap. By incorporating player statistics, shot charts, and court coverage data, this research shows that analytics are being used not only on the court but also in front office operations.

Another major theme across the literature is the league-wide shift in shot selection strategy. Zając et al. (2023) and Kilcoyne (2020) both provide evidence of the increasing preference for three-point shots and attempts near the basket, while jump shots inside the three-point line, called mid-range shots, have declined. This trend, often described as the "three-point revolution," reflects a larger movement toward maximizing offensive efficiency. Their work emphasizes how teams have restructured offenses and developed players to take advantage of higher expected value shot locations.

In addition to strategic analysis, several researchers have applied machine learning to performance prediction. Wang (2023) applied logistic regression, support vector machines, random forests, and deep neural networks to predict NBA game outcomes based on team and player-level statistics. Although Wang's research focused on predicting wins rather than individual shot outcomes, the modeling techniques, feature selection, and evaluation methods are directly relevant to the approach used in this study.

More directly aligned with my research, Meehan (n.d.) explored individual shot success prediction using traditional classification models. His findings showed that boosting models outperformed others, with shot distance and defender proximity serving as key predictive features. However, a major limitation of his work is that it relied on only one season of data, which reduces the ability to generalize the model's effectiveness across different players and eras. Harmon, Ebrahimi, Lucey, and Klabjan (2021) offered a more advanced and innovative solution by converting five-second multiagent trajectories into image data processed by convolutional neural networks. This approach captured spatial and temporal interactions between players, although it introduced complexity that may be difficult for non-technical stakeholders to interpret.

While these studies have laid the foundation for basketball shot prediction, common limitations are evident across much of the literature. Most rely on relatively narrow datasets, often from a single season, which limit the ability of models to account for long-term trends or player development. Additionally, several studies present their findings using technical language or modeling approaches that may be inaccessible to coaches, analysts, or fans without a background in data science. This project addresses those gaps by using a dataset of over 4 million shot attempts from 20 NBA seasons. The robustness of this dataset should allow the models to better account for player skill and the improvement of league wide shooting skill throughout the years. I also strive for interpretability, aiming to produce a model that is both accurate and understandable. By incorporating game context and player-specific variables, the study seeks to reveal the underlying factors that influence shot success, offering both academic insight and real-world value for basketball decision-makers.

In summary, the existing literature has helped shape the field of basketball analytics, particularly in understanding the evolution of strategy and the application of machine learning to sports prediction. However, there is a clear need for models that are trained on broader datasets and designed to be interpretable by both experts and non-experts. This research builds upon that foundation and aims to create a practical, scalable solution for understanding what it takes to predict a shot being made in the NBA.

IV. Methods

The dataset used in this study is titled "NBA Shots" and was obtained from Kaggle, an open data platform widely used for data science and machine learning projects. Curated by the user *Mexwell*, the dataset includes detailed shot-level information from NBA regular-season games spanning 20 seasons, from 2003–2004 through 2023–2024. It contains over 4 million individual shot attempts, making it one of the most comprehensive publicly available datasets for basketball analytics. Each record includes a variety of features such as season, team ID, team name, player ID, player name, position group, position, game date, game ID, home team, away team, event type (made shot or missed shot), shot made (true or false), action type (jump shot, layup shot, driving layup, turnaround jump shot, running jump shot), shot type (two-point or three-point), basic zone (mid-range, restricted area, in the paint(non-restricted area), above the break 3, left corner 3), zone name, location of shot on the x axis, location of shot on y axis, shot distance (ft), quarter, minutes left in quarter, seconds left in quarter and whether the shot was made or missed. These variables allow for a rich analysis of both player-specific and contextual factors that influence shot success.

This dataset is valuable to my research because of the volume of shots it contains. As previously stated, existing studies on shot prediction are limited to a single season, which can make it difficult to generalize findings across different players, teams, and eras. In contrast, the inclusion of data from two decades of NBA seasons enables more robust modeling and supports long-term trend analysis. The data's granularity also allows for advanced feature engineering and the application of machine learning techniques that account for real-time game dynamics.

For this project, I conducted all analysis using the R programming language within the RStudio environment. A variety of packages were used to support data manipulation, visualization, and model development. These included tidyr, dplyr, and data.table for data cleaning and wrangling; ggplot2 for data visualization; caret, xgboost, and keras for building and evaluating machine learning models; recipes for preprocessing; and purrr for functional programming tasks.

The procedures for this research followed a structured data science pipeline, beginning with software setup and ending in model evaluation. The raw data was stored across individual comma-separated value (CSV) files, each representing a different NBA season from 2003–2004 through 2023–2024. These files were loaded and combined into a single data set. Initial data inspection involved checking variable types for consistency and identifying missing values. It was found that approximately 7,930 observations were missing values for the position and position group variables. Given the scale of the dataset, these observations represented a small fraction of the data and were removed to allow these variables to be included as reliable predictors. To prepare for modeling, the dataset underwent several cleaning steps. Based on prior knowledge, I knew that some teams had changed names and cities during this time frame, so I wanted to address that within the data. After ensuring that all changes in names and cities represented the same franchises, I decided that using the TEAM_ID variable would be the best way to go about using teams as predictors. Records for the Los Angeles Clippers had a few inconsistencies as well. I found "LA Clippers" as another team name within the dataset, so I changed all of those team name values to "Los Angeles Clippers" for the sake of continuity. Column names were standardized for clarity, such as renaming SHOT_DISTANCE to SHOT_DISTANCE_FT, and redundant variables such as POSITION, EVENT_TYPE, and ZONE_ABB were removed.

An exploratory data analysis (EDA) was conducted to better understand how certain variables impacted shot success. Visualizations revealed that there have been more missed shots than made shots over the past 20 seasons. The make-to-miss ratio for three-point attempts was approximately 25% lower than for two-point shots. Lastly, Figure 1 illustrates the U-shaped pattern in shot distance indicated that players tend to shoot either very close to or far from the basket. Analysis by quarter showed a slight decline in shot-making success as games progress, with notably fewer makes in the opening and closing minutes of each quarter.

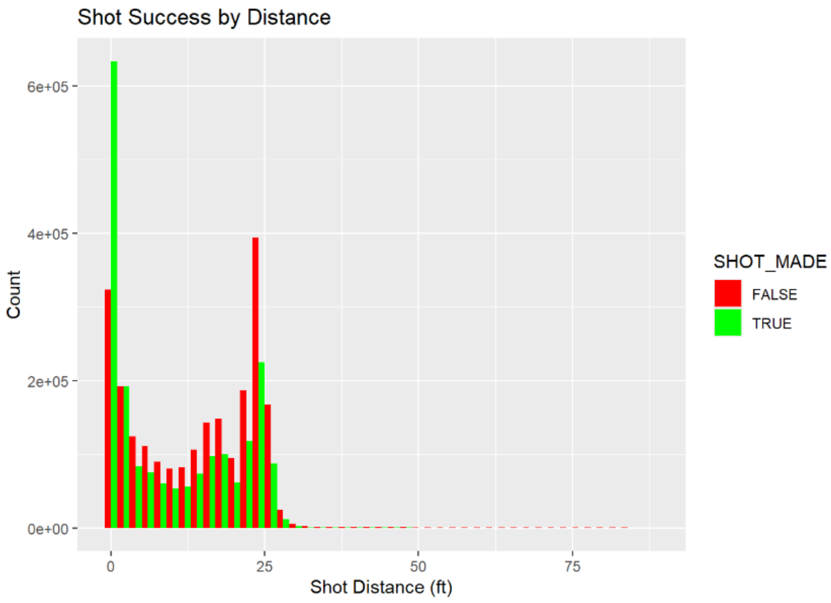


Figure 1: Graph showing the count of shots made and missed at each distance (ft) from the basket.

To enhance the dataset, several engineered features were introduced. A binary variable is_home was created to identify whether a shot was taken by the home or away team. This process involved mapping the abbreviations from the original HOME_TEAM variable to the correct team names using the TEAM_NAME variable. Once the correct abbreviations were matched to their respective team, I then compared the two variables. This resulted in a 1 if the team abbreviation matched the HOME_TEAM variable and a 0 if it didn't. Another variable, game_sec_elapsed, was engineered to represent the total seconds elapsed in a game at the moment of each shot. This was done by performing calculations on the QUARTER, MINS_LEFT, and SECS_LEFT variables. The target variable SHOT_MADE was also converted into a binary numeric form (1 for made, 0 for missed) to support classification modeling.

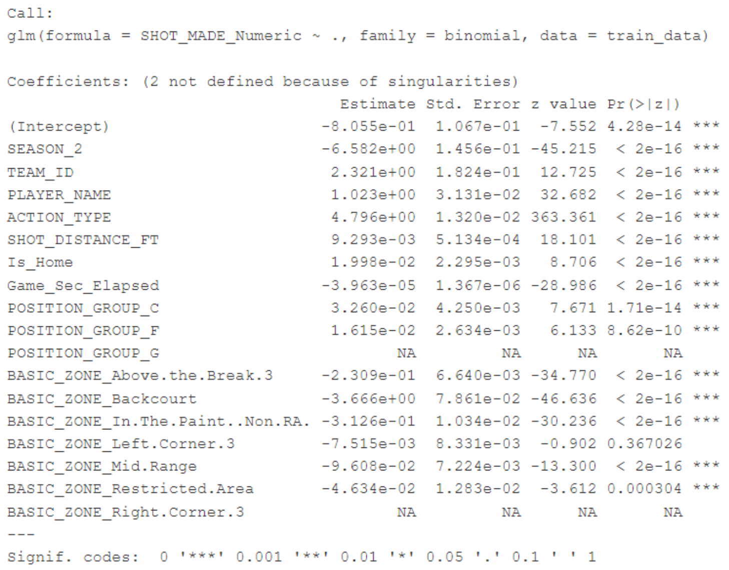
Next, variable relationships with the target variable were examined. A correlation matrix was used for numeric variables, and chi-square tests were applied to categorical variables. These are specific tests that compute just how impactful each variable is on the target variable. While most numeric variables showed weak correlations, variables like SHOT_DISTANCE_FT and LOC_Y had slightly stronger associations with shot success. Categorical variables such as ACTION_TYPE, SHOT_TYPE, BASIC_ZONE, and ZONE_NAME had particularly strong associations with the target, suggesting the location and type of shot significantly affects the outcome of a shot.

Using this analysis, relevant features were selected for modeling. These included SEASON_2, TEAM_ID, PLAYER_NAME, POSITION_GROUP, ACTION_TYPE, BASIC_ZONE, SHOT_DISTANCE_FT, IS_HOME, GAME_SEC_ELAPSED, and SHOT_MADE_NUMERIC. To prepare these variables, one-hot encoding was applied to POSITION_GROUP and BASIC_ZONE, while k-fold encoding was used for high-cardinality categorical variables like SEASON_2, TEAM_ID, PLAYER_NAME, and ACTION_TYPE. Feature selection and encoding are critical steps in building an effective machine learning model. The use of one-hot encoding for the variables POSITION_GROUP and BASIC_ZONE allows the model to differentiate between categorical groups without imposing an artificial numeric relationship between them. Meanwhile, k-fold encoding was applied to high-cardinality categorical variables to preserve valuable information while avoiding the dimensional explosion that one-hot encoding would cause. This approach helped the models learn from these features more efficiently, reduce overfitting, and maintain generalizability.

The dataset was then split into training and testing sets using an 80/20 split. This means that 80% of the data was used to train the model and then the other 20% of data was used to in the model to test how well the model performs. Two machine learning models were developed to compare the differences. One model was a logistic regression model and the other a gradient boosting model. Both models were evaluated using confusion matrices and standard performance metrics to assess their predictive accuracy and reliability. In addition, the importance of each variable was investigated for each model. The logistic model used the summary function for feature importance and the gradient boost model used the xgb.importance function.

V. Results

The logistic regression model achieved an overall accuracy of 63%. All variables included in the model were found to be statistically significant, with p-values less than 0.001, indicating that each contributed meaningfully to the model’s prediction of shot success. The most influential predictor, based on the coefficient estimates, was ACTION_TYPE, which captures the specific kind of shot taken—such as a layup, jump shot, or driving layup. This variable had the greatest impact on the predicted probability of a shot being made. Figure 2 illustrates the overall summary of the model. The model estimate is used to determine the impact a feature has on the outcome, with positive numbers increasing the likelihood of a make and negative numbers decreasing the likelihood of a make.



```
Call:
glm(formula = SHOT_MADE_Numeric ~ ., family = binomial, data = train_data)

Coefficients: (2 not defined because of singularities)
                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      -8.055e-01  1.067e-01  -7.552 4.28e-14 ***
SEASON_2                         -6.582e+00  1.456e-01 -45.215  < 2e-16 ***
TEAM_ID                           2.321e+00  1.824e-01  12.725  < 2e-16 ***
PLAYER_NAME                       1.023e+00  3.131e-02  32.682  < 2e-16 ***
ACTION_TYPE                       4.796e+00  1.320e-02 363.361  < 2e-16 ***
SHOT_DISTANCE_FT                  9.293e-03  5.134e-04  18.101  < 2e-16 ***
Is_Home                           1.998e-02  2.295e-03   8.706  < 2e-16 ***
Game_Sec_Elapsed                 -3.963e-05  1.367e-06 -28.986  < 2e-16 ***
POSITION_GROUP_C                  3.260e-02  4.250e-03   7.671 1.71e-14 ***
POSITION_GROUP_F                  1.615e-02  2.634e-03   6.133 8.62e-10 ***
POSITION_GROUP_G                         NA         NA      NA       NA    
BASIC_ZONE_Above.the.Break.3     -2.309e-01  6.640e-03 -34.770  < 2e-16 ***
BASIC_ZONE_Backcourt             -3.666e+00  7.861e-02 -46.636  < 2e-16 ***
BASIC_ZONE_In.The.Paint..Non.RA. -3.126e-01  1.034e-02 -30.236  < 2e-16 ***
BASIC_ZONE_Left.Corner.3         -7.515e-03  8.331e-03  -0.902 0.367026    
BASIC_ZONE_Mid.Range             -9.608e-02  7.224e-03 -13.300  < 2e-16 ***
BASIC_ZONE_Restricted.Area       -4.634e-02  1.283e-02  -3.612 0.000304 ***
BASIC_ZONE_Right.Corner.3                NA         NA      NA       NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2: Summary of logistic model used to show feature importance

The gradient boosting model performed slightly better, reaching an accuracy of 64%. As shown in Figure 3, ACTION_TYPE again emerged as the most important feature, followed by SHOT_DISTANCE_FT, SEASON_2, and PLAYER_NAME. This reinforces the idea that the nature of the shot and the distance from the basket are two of the most critical variables in determining shot success. This is intuitive and aligns well with basketball knowledge. The third and fourth most important features in the gradient boosting model were particularly interesting. SEASON_2, which represents the season in which the shot was taken, likely reflects changes in league-wide shooting efficiency over time, such as the growing emphasis on three-point shooting. PLAYER_NAME captured individual player tendencies and skill levels, confirming that the model was able to learn from specific player behavior. These results suggest that both evolving league trends and individual player skill are significant factors in predicting whether a shot will be successful.

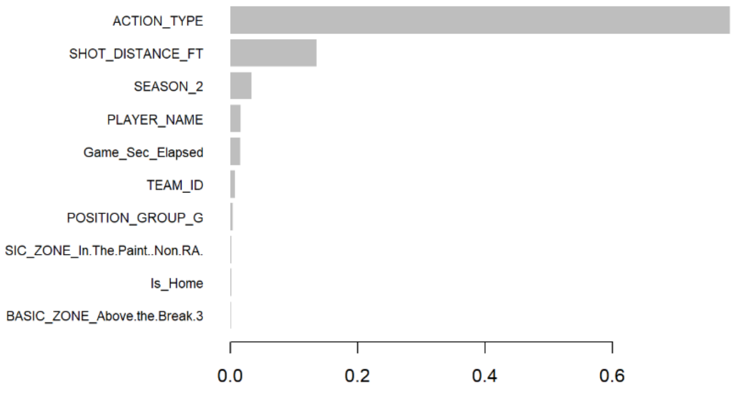


Figure 3: Feature importance of the gradient boosting model

VI. Discussion

The results of my study demonstrate both the complexity and the potential of using machine learning to predict NBA shot success. While the overall accuracy of the models, particularly the gradient boosting model at 64%, may not be sufficient for real-time implementation in coaching or decision systems, it does represent a meaningful improvement over a random guess. This outcome aligns with the primary goal of the research: to develop a model that performs better than chance while identifying the most influential factors associated with shot success.

One of the most valuable outcomes of this analysis was the ability to determine which features contributed most to model performance. Across both models, the type of shot (e.g., jump shot, layup, driving layup) emerged as the most important predictor. In the gradient boosting model, this feature held considerably more weight than any other, suggesting that the nature of the attempt heavily dictates the likelihood of success. Other highly ranked variables included shot distance, season, and player name. The importance of season may reflect broader league trends, such as the rising efficiency of three-point shooting in recent years. This highlights how shooting success has evolved over time since the early 2000s. The fact that player name was the fourth most important feature in the gradient boosting model suggests that the model successfully captured individual differences in shooting ability. This outcome supports the idea that the model learned from player-specific tendencies, and demonstrates the value of training on a large, multi-season dataset.

These findings build on and advance the existing literature. As shown in prior research, models that aim to predict shot success typically achieve similar accuracy levels. However, many of those studies relied on smaller datasets or focused narrowly on a single season. Like previous research, this study found that the gradient boosting model outperformed logistic regression in both accuracy and feature interpretation. What sets this work apart is the incorporation of data that spans two decades, allowing for greater insight into both individual player skill and broader changes in the game over time. The inclusion and importance of the PLAYER_NAME variable show that this model captured individual skill in a way that previous single-season models often could not.

The results of my research support my initial hypothesis: that it is possible to build a model that predicts shot success more accurately than random chance, and that key contextual and player-specific features can be identified. Not only did both models outperform a 50% baseline, but they also successfully highlighted interpretable, real-world variables that align with basketball knowledge and analytics.

Despite these encouraging outcomes, there are some limitations to acknowledge. Most notably, the dataset used in this study did not include defender proximity, which prior literature has shown to be an important factor in determining shot difficulty and outcome. Without accounting for defensive pressure, the model may overlook a critical aspect of in-game context. Additionally, the model did not incorporate real-time player movement or shot trajectory, which would likely further improve prediction accuracy. Future work could benefit from including player tracking data or SportVU features to build more nuanced models.

In summary, this research represents a step forward in the effort to develop interpretable, data-driven models for predicting shot outcomes in professional basketball. While there is still room for improvement, particularly in terms of feature richness and real-time inputs, the results show that machine learning can uncover meaningful patterns in shooting performance. This study offers a foundation for future work aiming to develop even more accurate and actionable models for use in basketball analytics.

VII. Conclusion

This study set out to develop machine learning models capable of predicting NBA shot success based on contextual and player-specific features. Using a comprehensive dataset spanning two decades and over 4 million shots, both a logistic regression model and a gradient boosting model were constructed and evaluated. The gradient boosting model performed slightly better with an accuracy of 64%, compared to 63% for the logistic regression model. While these accuracy levels suggest there are many unpredictable factors involved in shot outcomes, both models performed significantly better than a random guess, supporting the hypothesis and demonstrating the predictive potential of structured basketball data.

One of the most valuable findings from this research was the consistent importance of the ACTION_TYPE variable, which emerged as the most influential predictor in both models. This highlights that the nature of the shot is a key determinant of whether it is made. Additionally, shot distance, the season in which the shot occurred, and the player who took the shot were all identified as important variables. These results provide insight into not only individual player skill but also the broader evolution of shooting efficiency in the NBA over time.

The findings have broader implications for the field of basketball analytics. They show that models trained on large, multi-season datasets can capture meaningful patterns that go beyond basic shot location. By including features such as player identity and season, this research takes a step toward more individualized and era-aware performance analysis. While real-time use of such models still faces limitations, particularly due to missing features like defender proximity, this study provides a foundation for future efforts that aim to build more accurate, interpretable, and context-rich prediction systems.

VIII. Future Directions

The findings in this research suggest that using a large, multi-year dataset is essential for capturing player-specific context and uncovering deeper patterns in shooting performance. The inclusion of player identity as one of the most important features demonstrates the value of training models on long-term data, as it allows for a more complete understanding of individual tendencies and skill levels over time.

To build on this foundation, future research should incorporate defensive context, particularly defender proximity. Defense plays a critical role in shot success, and the absence of defensive pressure data in this study likely limited the model's predictive power. By pairing large-scale historical data with real-time or spatial defensive metrics, future models could provide a more complete and realistic picture of shot difficulty. This combination would lead to more accurate, context-aware models that better reflect the challenges players face on the court.

LIST OF REFERENCES

References

Harmon, M., Ebrahimi, A., Lucey, P., & Klabjan, D. (2021). *Predicting shot making in basketball learnt from adversarial multiagent trajectories*. arXiv. https://arxiv.org/abs/1609.04849v5

Kilcoyne, S. (2020). *The decline of the mid-range jump shot in basketball: A study of the impact of data analytics on shooting habits in the NBA* (Honors thesis, Bryant University). https://digitalcommons.bryant.edu/honors_mathematics/30

Meehan, B. (n.d.). *Predicting NBA shots* [Stanford University final project]. GitHub. https://github.com/BrettMeehan/CS229-Final-Project

Murakami-Moses, M. (n.d.). *Analysis of machine learning models predicting basketball shot success*. The American School in Japan.

Kambhamettu, A. R., Shrivastava, A., & Gwilliam, M. (2024). Quantifying NBA shot quality: A deep network approach. *Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports (MMSports '24)* (pp. 91–95). Association for Computing Machinery. https://doi.org/10.1145/3689061.3689068

Sliz, B. A. (2016). *An investigation of three-point shooting through an analysis of NBA player tracking data* (Master's thesis, Northwestern University). https://www.proquest.com/openview/5c3cbf83cf345eb71158a7ff2c5b4e2c/1

Siddique, S. (2024). Teaming strategy optimization: An analysis of NBA statistics, shot charts, and constraints (Master's thesis, Prairie View A&M University). https://digitalcommons.pvamu.edu/pvamu-theses/1

Wang, J. (2023). Predictive analysis of NBA game outcomes through machine learning. *Proceedings of the 6th International Conference on Machine Learning and Machine Intelligence (MLMI 2023)*. https://doi.org/10.1145/3635638.3635646

Zając, T., Mikołajec, K., Chmura, P., Konefał, M., Krzysztofik, M., & Makar, P. (2023). Long-Term Trends in Shooting Performance in the NBA: An Analysis of Two- and Three-Point Shooting across 40 Consecutive Seasons. *International Journal of Environmental Research and Public Health*, *20*(3), 1924. https://doi.org/10.3390/ijerph20031924

LINK TO THE REFERENCES