Analysis of NBA Player Offensive Efficiency using Artificial Intelligence

John Baldi, Shweta Madhale, Babafemi Sorinolu, Su Zhang
Stevens Institute of Technology,
Hoboken, New Jersey, 07303
USA

Abstract

We are looking to improve the understanding of a player's offensive efficiency by accounting for new, innovative ideas about the context behind the shots that a player takes throughout the game. Specifically, we plan to engineer features and apply artificial intelligence techniques to model on concepts such as player volume, player movement, shot openness, game time remaining, possession time remaining, the strength of the defense, and more. With these new features being accounted for, we believe we can provide a more accurate representation of a player's true offensive efficiency. A better understanding of player efficiency will be useful for NBA teams to prepare their gameplan for upcoming seasons.

Introduction

This project finds application in the sports domain. It focuses on analyzing player efficiency for building a successful team. Sporting events generate millions of data points, encompassing player actions, ball movements, team positioning, and event outcomes. Sports organizations are constantly trying to leverage this information to gain insights into the key factors that contribute to the success of individual players and teams. This is especially true for the National Basketball Association (NBA), as these organizations invest millions into research and development in an attempt to gain a competitive advantage.

The goal of this research is to improve an outdated idea that has been around since basketball was invented: "Player offensive efficiency". Historically, the initial approach to determining a player's offensive efficiency was to look at field goal percentages (FG%) for all shots, three-pointers (FG3%), and free throws (FT%), which merely divide the number of shots made by the number of attempts. Later, in 2002, statistician Dean Oliver developed "Effective field goal percentage" (eFG%), which is a weighted formula adjusting for the fact that a 3-pointer is worth more than a 2-pointer[1]. While eFG% provides a better idea of a player's shooting efficiency, it is becoming an increasingly naive approach when considering the vast quantity of data that is now tracked during NBA games.

Related Work

Several previous works are inspirations for, and included as references in, this project. The papers or studies were published from 2016 to present, which coincides with the development stage of the techniques of analysis of player efficiency, and they have a great impact on how we choose our approach to tackle the problem.

- Traditional measurement for players' efficiency treat performance in exactly the same way without considering game clock and game time which could decide the game outcome. Sameer and Shane proposed a win probability framework using Bayesian linear regression model to estimate an individual player's impact on the court[2]. And also introduce several posterior summaries to derive rank-ordering of players within their team and across the league.
- To overcome some of the limitations of the Ei (Efficiency Index) and GRS (Game-related statistics), José, Gilbert and Leonardo (2022) introduced the new index BEi (Basketball Efficiency Index) and BPi (Basketball Productivity Index) in their paper[3].The main purpose of this index is to summarize the players' performance in terms of both efficiency and productivity, as well as overcome the bias derived from both the players' playing time and game pace.
- Rodolfo and Giorgio, in their paper, evaluate each player's importance and the player's marginal contribution to the utility of an ordered subset of players, through a generalized version of Shapley value, to determine the probability a certain lineup has to win the game[4].
- Using Multistep Markov chains and Bayesian models, in his paper, Joshua (2017) focuses on a certain player's (Lebron James) free throw through 2016-2017 to determine whether a player has a hot hand[5].
- To provide suggestion for the coaches and management team during NBA draft, Adarsh, Brian, Brandon and Sohail (2018) utilize random forest classifier, logistic regression and support vector machine to determine the most relative biometric data for the success of a college player, and find despite of the players' performance, wingspan, height, and reaches were the most important[6].

Implementation Details

The implementation of our system was done using Python and several artifical intelligence tools and techniques.

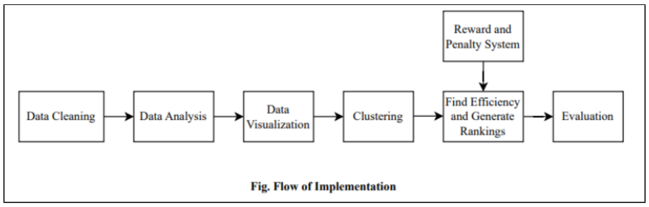

Reward and Penalty System
Data Cleaning
Data Analysis
Data Visualization
Clustering
Find Efficiency and Generate Rankings
Evaluation
Fig. Flow of Implementation

Figure 1: System Details

Data

The data set was obtained from Kaggle and containes individual shot attempt data for the 2014-2015 NBA season. The raw dataset contains 21 features about the context and results of the attempts such as Game ID, Matchup, Location, Final Margin, Shot Number, Period, Game Clock, Shot Clock, Dribbles, Touch Time, Shot Distance, Points Type, Closest Defender, Closest Defender ID, FGM, Player Name, Player ID. The raw sample size is 128,070 shots.

Data Cleaning

Before any artificial intelligence techniques could be applied, the data needed to be prepared. First, irrelevant features like Matchup, Location, W, Shot Result, and Points Type, were dropped and not used further. These features do not give us any information about the in-game context of the shot itself. Next, several values of the touch time feature are illogically negative, so the shots attempts with these values were dropped. Furthermore, a feature for seconds left in the period was engineered from the game clock feature in format which is in the format 'mm:ss'.

Data Analysis

The raw features were resourceful but did not fully capture the context of the shots being taken by themselves, so we performed data analysis to extract more information for the final solution. An additional feature we engineered was finding a defender's blocking efficiency, which was a ratio obtained by count of the field goals blocked by the defender to the total number of shots attempted against the defender.

$$BlockingEfficiency = \frac{Count(FGMAgainst = 0)}{Count(FGMAgainst)} \tag{1}$$

Using this feature, we could create a rank of the best defenders, and understand how shooting against a better defender may be affecting a player's shooting efficiency.

Data Visualization

Next, we visualized our features against the shot outcome to understand their relationships. In the visualizations, 1 denotes a made shot, and a 0 is a miss. Through these plots, we were able to gather which variables may have an impact on the outcome of a shot, and we selected these features to move forward with in the application of our AI techniques.

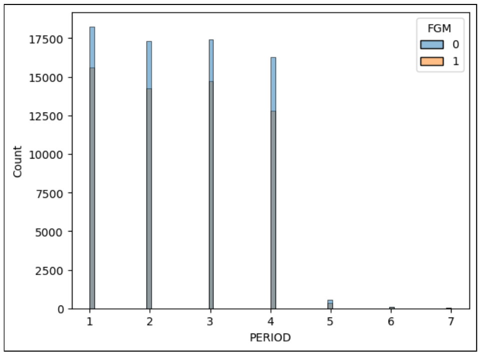


Figure 2: Histogram of Period vs FGM

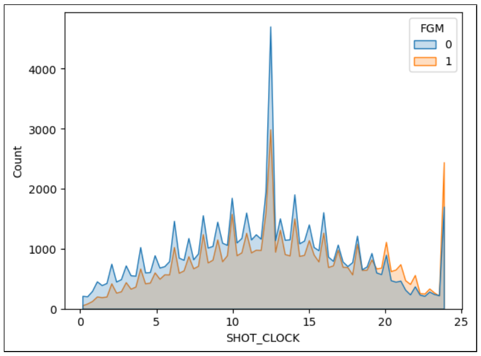


Figure 3: Histogram of Shot Clock vs FGM

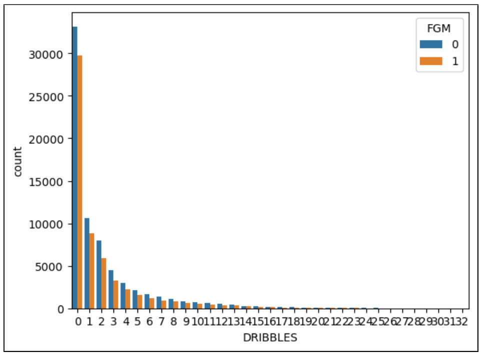


Figure 4: Count plot of Dribbles vs FGM

Clustering

Our approach includes using a clustering algorithm to group the shots attempted by the features that affect the shot outcome. We ultimately chose the K-Means algorithm because we are working with big data and several explanatory features. Relating this to the problem statement, the idea is to group the shots by their difficulty, to essentially standardize the shots attempted within each cluster.

To optimally decide the number of clusters, we chose to explore the Silhouette Method. The Silhouette Method scores the k-value chosen based on its similarity or cohesion within clusters against its separation between clusters. We saw this as an applicable method as we are trying to sep-

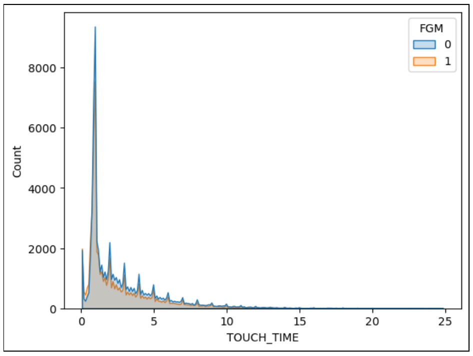


Figure 5: Histogram of Touch Time vs FGM

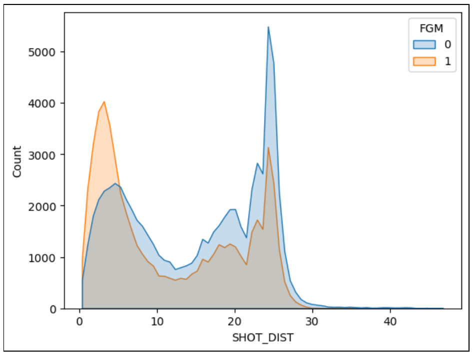


Figure 6: Histogram of Shot Distance vs FGM

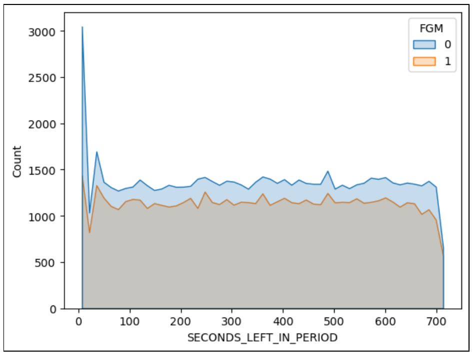


Figure 7: Histogram of Seconds Left vs FGM

arate and standardize the context behind different shots that are attempted in NBA games.

As seen in the Figure 10, the optimal k-value for our model given our features is 4.

Next, we implemented our clustering model with k=4.

In Figure 11, we plot the clusters created by the model. While all of our features were used in the clustering model, we cannot view all of them in one visualization, so we plotted the clusters with shot distance and closest defender distance as these features have a major effect on shot outcome.

To explore whether the clustering model was successful in grouping the shots by difficulty, we looked at the shooting efficiency within each cluster. Shooting efficiency is simply

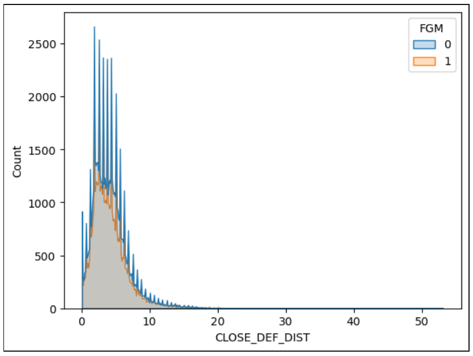


Figure 8: Histogram of Defender Distance vs FGM

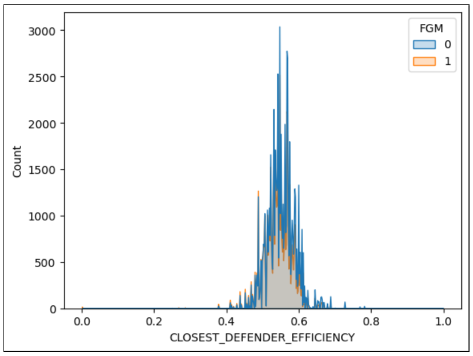


Figure 9: Histogram of Defender Efficiency vs FGM

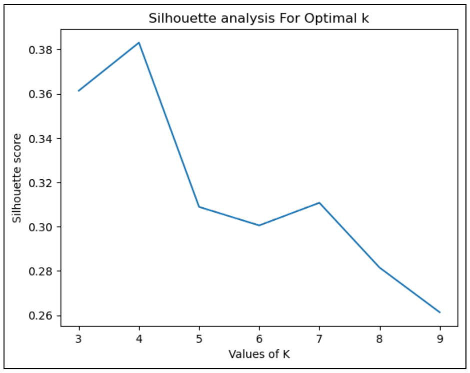


Figure 10: Silhouette Scoring for Optimal k

the number of made attempts divided by the total number of attempts.

$$ShootingEfficiency = \frac{Count(FGM = 1)}{Count(FGM)} \qquad (2)$$

In Table 1, we see the shooting efficiencies by cluster, and we observe a significant difference between them. We can tell that, on average, shots in cluster 2 are made the most often, and hence can be considered easiest amongst the clusters, while the shots in cluster 0 are the most difficult. These clusters give us the standardized grouping needed for us to better determine which players are actually more efficient.

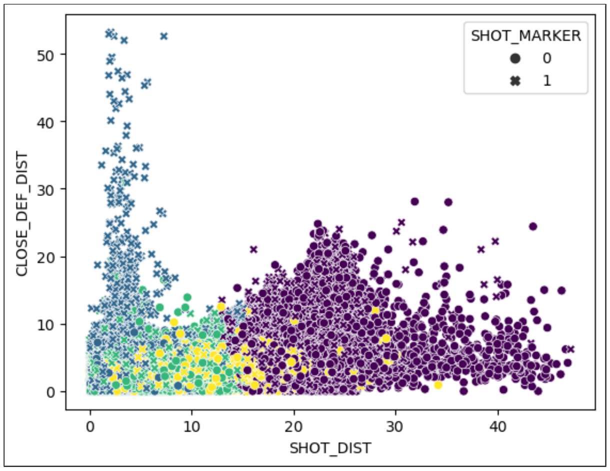


Figure 11: Clusters

Cluster No	Shooting Efficiency
0	0.37379
1	0.60036
2	0.48430
3	0.41179

Table 1: Shooting Efficiency by Cluster

Reward and Penalty System

Next, we developed the efficiency reward and penalty system. The goal within this system was to reward players for making more difficult shots, and penalize them for missing easier shots, with the difficulty of the shots being derived from the clustering.

To do this, we compared the shooting efficiency within each cluster to the total average shooting efficiency of the data, and produced equations for our Reward Factor and Penalty Factor.

$$Reward = 2(ClusterEfficiency - AverageEfficiency) \tag{3}$$

$$Penalty = 2(AverageEfficiency - ClusterEfficiency) \tag{4}$$

The total average shooting efficiency was about 0.46, so with this system in place, players would be rewarded for shots taken within clusters 0 and 1 and penalized for shots taken within clusters 2 and 3, with clusters 0 and 2 being more heavily weighted.

These factors are added or subtracted, depending on reward or penalty, from the count of each player's shot make and attempt to better determine the overall efficiency.

Evaluation

Using the clusters to standardize shot difficulty and the reward and penalty system, we were able to calculate each individual player's adjusted shooting efficiency. To evaluate our results, we compared our efficiencies and our efficiency rankings to the rankings from using an efficiency metric currently used by teams.

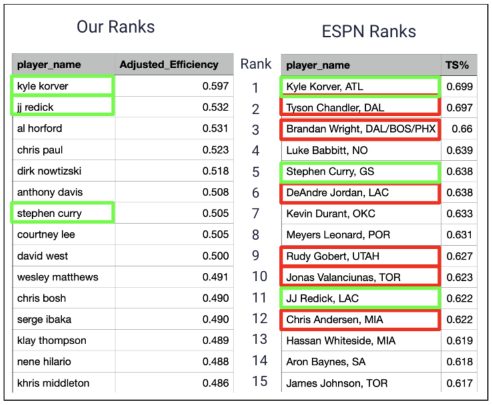
We gathered data from the 2014-2015 season from ESPN and retrieved a list of players with the highest True Shooting Percentage (TS%), which is efficiency adjusted for the type of shot taken (free throw vs 2-point vs 3-point).

To compare our rankings with ESPN's, we used Rank-Biased Overlap (RBO). RBO provides a similarity score between two lists.

$$RBO(S,T,p) = (1-p)\sum_{d=1}^{\infty}(p^{d-1}).A^{d} \tag{5}$$

Our RBO score came out to 58.084%. This score demonstrates that similarity does exist between our metric and current metrics, but there is also significant difference between the lists, so our metric can bring new insights to which players are more efficient.

Furthermore, when looking at the most efficient players, both by our adjusted metric and ESPN's TS%, intuitively it's clear that our implementation was successful in accomplishing our initial goal of context-adjusted efficiency.



Our Ranks

ESPN Ranks

player_name	Adjusted_Efficiency	Rank	player_name	TS%
kyle korver	0.597	1	Kyle Korver, ATL	0.699
jj redick	0.532	2	Tyson Chandler, DAL	0.697
al horford	0.531	3	Brandan Wright, DAL/BOS/PHX	0.66
chris paul	0.523	4	Luke Babbitt, NO	0.639
dirk nowtizski	0.518	5	Stephen Curry, GS	0.638
anthony davis	0.508	6	DeAndre Jordan, LAC	0.638
stephen curry	0.505	7	Kevin Durant, OKC	0.633
courtney lee	0.505	8	Meyers Leonard, POR	0.631
david west	0.500	9	Rudy Gobert, UTAH	0.627
wesley matthews	0.491	10	Jonas Valanciunas, TOR	0.623
chris bosh	0.490	11	JJ Redick, LAC	0.622
serge ibaka	0.490	12	Chris Andersen, MIA	0.622
klay thompson	0.489	13	Hassan Whiteside, MIA	0.619
nene hilario	0.488	14	Aron Baynes, SA	0.618
khris middleton	0.486	15	James Johnson, TOR	0.617

Figure 12: Top-Ranked Players by Efficiency Metrics

In Figure 12, you can see the top-ranked players by our efficiency metric and ESPN's. Our list, along with ESPN's, includes players consistently taking more difficult shots (highlighted in green), but doesn't include players consistently taking easier shots(highlighted in red).

Conclusion

Our approach and implementation was successful in accomplishing our initial goal. We were able to evaluate the context around NBA shot attempts, and then apply artificial intelligence techniques on that context to develop a new way of viewing offensive player efficiency.

There is currently no metric that adjusts for the type of shot context we considered in this research. We ultimately believe NBA teams and coaches could use our metric, or similar approach, to determine which players are more efficient and gain a competetive advantage.

Future Scope

While our implementation was successful, a lot can be done to improve and continue this research. First, better and more current data can be used. The data we used had several flaws and was from 2014-2015, so it can't necessarily be applied in a business sense to today's players. Furthermore, with more detailed data around the context of the shot (player movement, shot angle, defensive positioning), even more accurate metrics can be produced.

References

[1] Kubatko, Justin & Oliver, Dean & Pelton, Kevin & Rosenbaum, Dan. (2007). A Starting Point for Analyzing Basketball Statistics. Journal of Quantitative Analysis in Sports. 3. 1-1. 10.2202/1559-0410.1070.

[2] Deshpande, Sameer K. and Jensen, Shane T.. "Estimating an NBA player's impact on his team's chances of winning" Journal of Quantitative Analysis in Sports, vol. 12, no. 2, 2016, pp. 51-72. 2015.

[3] Senatore, J., Fellingham, G., & Lamas, L. "Efficiency and productivity evaluation of basketball players' performance.", Motriz- Revista de Educação Física, 2022, doi- 10.1590/s1980-657420220004922

[4] Metulini, R., Gnecco, G. Measuring players' importance in basketball using the generalized Shapley value. Ann Oper Res (2022). https://doi.org/10.1007/s10479-022-04653-z

[5] Chang, Joshua C.. "Predictive Bayesian selection of multistep Markov chains, applied to the detection of the hot hand and other statistical dependencies in free throws." Royal Society Open Science 6 (2017): n. Pag.

[6] Adarsh Kannan, Brian Kolovich, Brandon Lawrence, Sohail Rafiqi. "Predicting National Basketball Association Success: A Machine Learning Approach". 2018.

[7] Inan, Tugbay, and Levent Cavas. "Estimation of Market Values of Football Players through Artificial Neural Network: A Model Study from the Turkish Super League." Applied Artificial Intelligence 35, no. 13 (2021): 1022–42. https://doi.org/10.1080/08839514.2021.1966884.

[8] Yang, Zhuo. "Research on Basketball Players' Training Strategy Based on Artificial Intelligence Technology." Journal of Physics: Conference Series 1648, no. 4 (2020): 042057. https://doi.org/10.1088/1742-6596/1648/4/042057.

[9] García-Aliaga, Abraham, Moisés Marquina, Javier Coterón, Asier Rodríguez-González, and Sergio Luengo-Sánchez. "In-Game Behaviour Analysis of Football Players Using Machine Learning Techniques Based on Player Statistics." International Journal of Sports Science & Coaching 16, no. 1 (2020): 148–57. https://doi.org/10.1177/1747954120959762.

[10] Beal, Ryan, Timothy J. Norman, and Sarvapali D. Ramchurn. "Artificial Intelligence for Team Sports: A Survey." The Knowledge Engineering Review 34 (2019). doi.org/10.1017/s0269888919000225.

[11] Li, Bin, and Xinyang Xu. "Application of Artificial Intelligence in Basketball Sport." Journal of Education, Health and Sport 11, no. 7 (2021): 54–67. doi.org/10.12775/jehs.2021.11.07.005.