**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | Capstone Project Proposal - Machine Learning for NBA Players Performance Analysis |
| **Lecturer Name:** | James Garza |
| **Student Full Name:** | Sertaç Battal |
| **Student Number:** | 2023472 |
| **Assessment Due Date:** | 17/12/2023 |
| **Date of Submission:** | 17/12/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Introduction**

Together, we held physical meetings in and outside of the college environment to discuss the topic we were assigned and researched datasets about topics that we both have a common interest in. As a result of these meetings, we decided that we will continue our work on our joint decision, the National Basketball Association (NBA).

The National Basketball Association is known worldwide as one of the most popular

and prestigious leagues in basketball (Jozsa, 2011). The NBA presents an exciting

sporting event filled with talented players, highly variable scores, a wide range of statistics, and average of more than 1.5 million viewers per game (Paulsen, 2023) in the 22/23 season just in the United States.

With a total league revenue of over 10 billion dollars during the 2020/21 season (Statista Research Department, 2023), revenue streams ranging from television contracts and sponsorships to merchandise and tickets for the games, some teams achieved valuations of over $4 billion (Forbes, 2020) and revenues compared to the biggest football clubs in the world, football being the most watched sport worldwide (Deloitte, 2022).

Just these facts could be enough to justify the choice of NBA as the subject of this report, but our passion for the sport and the huge amount of data available for study and research made available by the NBA and other sources makes it even more important and interesting, with simple statistics like game results, rebounds and minutes played by player available since the 1950’s and more advanced data like shot charts made available since the early 2000’s.

This report will help us understand the transformation of the NBA over the past years and make sense of our assessment of potential future success trends in the league. Data analysis and statistical results will be an important reference source for NBA and basketball lovers.

**Objectives**

The objective of this capstone project is to gather, explore and analyse the great amount of data available about NBA games and players and obtain useful information like which teams have the highest chance of being the champion, which ones present the biggest probability of reaching the playoffs, the importance of certain players contributing to achieve the playoffs when compared to other team members, and the impact that players coming from the bench can make during the season.

Succeeding to achieve the goals written above can allow sponsors to better define the amount of money that can be spent on sponsorships, teams to respond to it according to their performance forecast and better evaluate their television deals to appear on more prime time games, attracting more money from sponsors, bringing the opportunity to earn more money from merchandise and maybe influencing more young prospects to join the organisation.

**Scope**

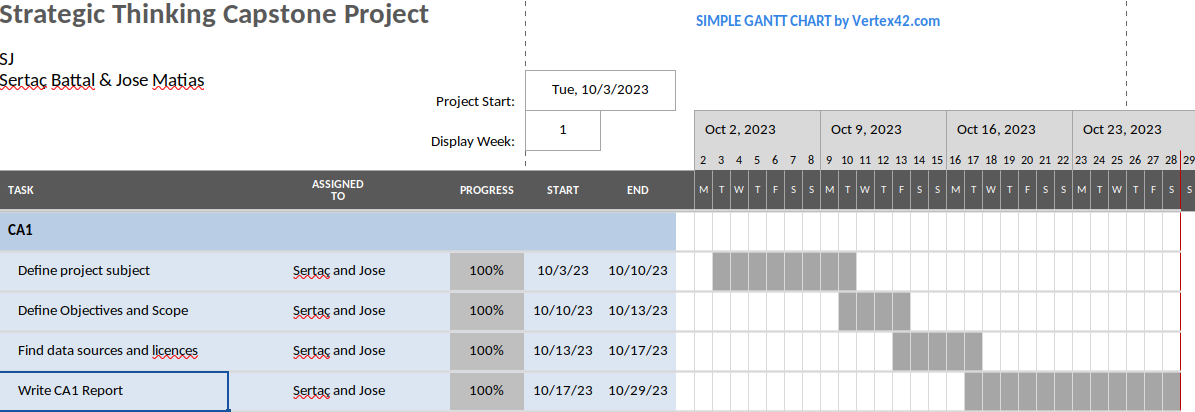
In this two-semester graduation project, the necessary data for NBA player performance and team achievements will be examined and a study will be conducted for in-depth analysis on this topic. The data to be used in the project is planned to be created by taking from the API provided on the NBA website. For this purpose, the API data retrieval technique will be used in the project. In addition, together with my team colleagues, we preferred to apply CRISP-DM (Cross-Industry Standard Process for Data Mining) for the management of this project. The following components will be included in the Capstone project:

**Player performance analysis**: Player productivity will be measured in the analysis to be performed with the scoring trends, shooting accuracy, individual player scoring, assists and rebounds data found in the data we have obtained through the NBA api. As a result of these analyses, insights about the achievements between the player's individual performance and team success can be obtained.

**Team Success Factors**: An in-depth study of the factors affecting team success will be conducted. This analysis aims to cover team performance, player achievements, coaching strategies and their effects on overall team success.

With the aim of achieving these goals, the project will focus on learning the techniques of data retrieval from the API and adhering to the CRISP-DM methodology, which covers business understanding, data understanding, data preparation, modelling and evaluation.

In order to ensure efficiency in the project, by adhering to the CRISP-DM laws, the tasks to be done in the project will be divided into different stages. In order to facilitate visual management and ease of making updates of the project status, we will use a GANTT diagram. Our GANTT chart is shown in Figure 1.



*Figure 1: Gantt chart representing the first phase of the capstone project.*

**Data Source and Ethical Considerations**

In relation to the data sources, there are plenty of websites providing data related to the NBA, including game statistics, athletes biometric information, fan engagement and

attendance to the games, financial figures and so on. Despite the volume of data available, most of these websites do not allow downloading the data for commercial and academic purposes for free, because of the effort they have made to present clean and organised, sometimes even tailor made datasets available from the raw data available online, and it is natural for them to make profit based on their work.

The main source of data for this report is going to be the NBA Stats website, which states that “By using such NBA Statistics, you agree that: (i) any use, display, or publication of the NBA Statistics shall include a prominent attribution to NBA.com in connection with such use, display, or publication; (ii) the NBA Statistics may only be used, displayed, or published for legitimate news reporting or private, non-commercial purposes”. The full disclaimer of the use of data from NBA Statistics can be found on https://www.nba.com/termsofuse#nba-statistics.

To have access to the data and further analyse it, we will get datasets from Kaggle that extracted the information from NBA Statistics and also utilise APIs to get any information that we need in order to feed the model and achieve the objectives of this project. All this information about the game results and players information present no ethical concerns, since the games are televised and the biometric data about the players is widely available, with no sensitive data being used.

**Project Goals and Objectives**

This project aims to analyse the performance of NBA players. In particular, it is aimed to understand the impact of players' statistics on team success and the relationships between these statistics.

Additionally, while examining the data in the dataset, the following questions were answered in the statistics we looked at first.

Which player stats are correlated with each other?

How are minutes played distributed?

How has the game changed over the past 5 years?

And Comparing Regular Season to Playoffs

**Data Collection and Data Cleaning**

The project includes 5-year player statistics for the NBA 2018-2022 seasons. The data set includes the basic performances of the players. The data used was obtained from the API provided to us on the NBA’s website. The analysis was continued by first taking the data from the API, converting it into a csv file, and then reading it from the created ‘nba\_player\_data.csv’ file. The data sources to be used for the project were determined and these data were cleaned, missing values ​​were eliminated and pre-processing steps were completed. Column descriptions of the data set used are as follows;

**Year**: The year in which the player’s performance was recorded.

**Season\_type**: The type of the season (for example, “Regular Season” or “Playoffs”).

**PLAYER\_ID**: The player’s unique identification number.

**RANK**: Player’s rank.

**PLAYER**: Name of the player.

**TEAM\_ID**: The unique identification number of the team.

**TEAM**: The team the player plays for.

**GP**: Number of games played (Games Played).

**MIN**: The player’s total time taken in minutes.

**FGM**: Number of successful field goals (Field Goals Made).

**FGA**: Total number of field goals (Field Goals Attempted).

**FG\_PCT**: Field Goal Percentage.

**FG3M**: Number of successful three-point field goals (Three-Point Field Goals Made).

**FG3A**: Total number of three-point field goals (Three-Point Field Goals Attempted).

**FG3\_PCT**: Three-Point Field Goal Percentage.

**FTM**: Number of successful free throws (Free Throws Made).

**FTA**: Total number of free throws (Free Throws Attempted).

**FT\_PCT**: Free Throw Percentage.

**OREB**: Number of offensive rebounds collected (Offensive Rebounds).

**DREB**: Number of defensive rebounds collected (Defensive Rebounds).

**REB**: Total number of rebounds (Total Rebounds).

**AST**: Number of assists (Assists).

**STL**: Number of steals (Steals).

**BLK**: Number of blocks made (Blocks).

**TOV**: Number of turnovers (Turnovers).

**PF**: Number of fouls committed (Personal Fouls).

**PTS**: Total points scored (Points).

**EFF**: Player’s efficiency value (Efficiency).

**AST\_TOV**: Assist to Turnover Ratio.

**STL\_TOV**: Steal to Turnover Ratio.

Additionally, we created new columns in our statistical analysis and the explanation of the columns we used is as follows;

**FG% (Field Goal Percentage):** The player’s on-court shooting accuracy. It is obtained by dividing the total successfully completed on-court field goals by the total on-court field goal attempts.

**3PT% (Three-Point Percentage):** The player’s accuracy in three-point shooting. It is obtained by dividing the total successfully completed three-pointers by the total three-point attempts.

**FT% (Free Throw Percentage):** The player’s accuracy in free throw shooting. It is obtained by dividing the total successfully completed free throws by the total free throw attempts.

**FG3A% (Three-Point Attempt Rate):** The ratio of three-point attempts to the total on-court field goal attempts.

**PTS/FGA (Points per Field Goal Attempt):** The number of points a player scores per attempted field goal.

**FG3M/FGM (Three-Pointers Made per Field Goals Made):** The ratio of made three-pointers to made on-court field goals.

**FTA/FGA (Free Throw Attempt Rate):** The ratio of free throw attempts to the total on-court field goal attempts.

**TRU% (True Shooting Percentage):** The player’s true shooting accuracy from total on-court field goals, considering free throws. It is calculated with the inclusion of free throws.

The features of the columns we chose during data cleaning were in line with the interests stated in the purpose of the project.

**Data Analysis and Exploratory Data Analysis (EDA)**

With EDA steps, basic analyzes were carried out on player statistics and team performance. Data exploration was done using histograms, scatter plots and scatter plots. The results obtained in response to the questions stated in the purpose of the project are as follows.

The columns we used in these analyses;

'MIN','FGM', 'FGA','FG3M','FG3A','FTM','FTA','OREB','DREB','REB','AST','STL','BLK', 'TOV','PF', 'PTS'

**Which player stats are correlated with each other?**

In this case, the aim is to understand whether different features or variables in the athletes' statistics data set are related to each other.

For example, you might want to understand the correlation between a basketball player's shooting percentage (FG%) and the number of three-pointers he makes. If there is a positive correlation between these two features, that is, as the player's three-point percentage increases, the number of three-pointers he makes also increases, we can say that there is a positive relationship between these two features.

This type of analysis can be used to understand how a player's particular skills or statistics relate to each other, to evaluate team performance, or to assist in player selection.

First of all, when we look at the correlation between the data, it gives us the graph below.metin, ekran görüntüsü, renklilik, kalıp, desen, düzen içeren bir resim

Açıklama otomatik olarak oluşturuldu

If we are going to examine the correlation values in these five seasons, we can give a few examples in this way.

FG3M - OREB (Three-Pointers Made - Offensive Rebound): -0.66

There is usually an inverse relationship between the number of threes a player scores (FG3M) and the offensive rebound he receives (Dec). This also means water, indicating that players who score more points with their ucluk shots may be less effective in hucum rebounds.

FGM - FGA (Field Goals Made - Field Goals Attempted): 0.88

There is a strong positive relationship between the number of on-court baskets scored by a player (FGM) and the total number of on-court shots (FGA). Dec. The more on-court baskets a player throws, the more often the total number of on-court shots also increases. In other words, the players in the five seasons we examined in the data set contributed to the score with low scores.

REB - FG3A% (Rebounds - Three-Point Attempt Rate): -0.61

There is a relationship between a player's three-Decker shooting rate (FG3A%) and the total number of rebounds (REB), which usually move against each other. As the tip shooting rate increases, the total number of rebounds usually decreases.

FG% - TRU% (Field Goal Percentage - True Shooting Percentage): 0.81

There is a strong positive relationship between the on-field hit rate (FG%) and the actual hit rate (TRU%). Dec. This usually indicates that as the on-field hit rate increases, the actual hit rate also increases. This strengthens the correlation relationship between FGM - FGA.

Note:

Positive Correlation: There is some type of linear relationship between two variables. When one variable increases, the other variable usually increases as well.

Negative Correlation: There is an inverse relationship between two variables. When one variable increases, the other variable usually decreases.

**How are minutes played distributed?**

The graph below shows that the players are playing in a few minutes interval per hundred. For example, between 0-99 minutes, 14% of players in these five seasons play, while this ratio decreases to 0.3 in the minutes between 2800 and 2899.

diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

At least 200 minutes of players remaining in the regular season and playing 5 matches,

and when we look at the graphs simultaneously for the players who stayed for at least 5 minutes and played at least 1 match for the playoffs, we get a graph as follows

ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, piksel içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Conclusion**

If we look at the middle section on the graph, the total players' stay in the game between the 12th and 34th minutes is 0.82 in the regular season, while the stay in the game in the playoffs is 0.47. This shows us that there is more player circulation in the playoffs, macs are moving faster.

**How has the game changed over the past 5 years?**

In this process is used to standardize the statistics of players per minute and compare players with different playing times. It can scale data by calculating players' ratings per minute. For this, we proportion our data by considering that a match is 48 minutes.

metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

In order to evaluate the players under similar conditions, we can perform calculations as follows to find the score opportunities obtained per 100 positions.

öykü gelişim çizgisi; kumpas; grafiğini çıkarma, metin, çizgi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Conclusion**

For 'Fig 1' Graph

When the individual columns in the graph shown in the table above are examined, there have been decreases or increases over the years. For example, TRU% (True Shooting Percentage) draws a higher graph. The BLK(Block) value, that is, a drop has been observed in the blocks since 2018. In the same way, REB, i.e. rebunda, has been falling for years. Here, the COVID period shows a serious change in the data, especially if there are no data in the 2021 season, a deviation in the rate may be seen.

For 'Fig 2' Graph

This graph shows how the changing statistics (player performance) have evolved over the seasons. According to the values given in the graph above, water comments can be made. For example; DREB (Defensive Rebounds) has fallen over the years in hindsight, this may indicate a physical change of the player, a change in playing style or a change in team tactics. On the other hand The decrease of DREB and the increase of OREB may indicate that the player's effectiveness on the offensive side has increased or that his focus on defensive rebounds has changed. If we look at it numerically, 3PT%(Three-Point Percentage), that is, the value of the three-number scores is considered as the peak point in 2020, except for this, with the pandemic period together they are looking at the value of min in 2021. If we look at the TRU%(True Shooting Percentage), the rate continues to increase even if it falls in 2021.

**Comparing Regular Season to Playoffs**

First of all, by dividing our data set into two separate data frames as regular season and playoffs season, we can find the effects of playoffs on player data compared to the regular season by comparing these two seasons within the season years.

The graphs of the new values ​​formed after the necessary calculations are made are as follows.

metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, metin, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi, diyagram, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Conclusion**

The graph values above give us the ratios of player performances in the playoffs compared to the regular season. If we are going to interpret the value from the graph by liking; for example, the decrease in the value of FTA (Free Throw Attempts) over the years We can say that the decrease in free throw attempts by players may be due to changes in the way players play, the strategies of teams, or a general trend in the league. The value of DREB (Defensive Rebounds) decreases until 2021 and increases in 2022, we can interpret that the players/teams have switched to a defensive tactic in the last year and have become defensive in the last season, while they have been more competitive in the playoffs compared to regular seasons before. When we look at the AST\_TOV(Assist to Turnover Ratio) value, it seems that a player's ability to manage the team's offense and make effective passes to teammates has increased in recent years, compared to the regular season in the playoffs. Finally, when the graph of the TRU% (True Shooting Percentage) value is examined, it can be seen that the shooting rate of the players in the last year has decreased compared to the regular season in the playoffs.

**Business understanding**

It aims to predict a player’s impact on the score by using NBA players’ data.

What characteristics among player statistics can increase or decrease a team's chances of success? (Based on players who scored more baskets than average, according to player data over 5 seasons)

**Data Cleaning and Visualization**

First of all, the Year, Season\_type, PLAYER\_ID, RANK, TEAM\_ID, TEAM columns, which were in our data set but meaningless for our model, were cleared. Afterwards, a graph was obtained by using the 'missingno' library to see the null values ​​in our data set. And as can be seen in the graph below, we do not have a null value.

dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Graphically examining the relationships between player data.

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, çizgi, renklilik içeren bir resim

Açıklama otomatik olarak oluşturuldu

As can be seen from the graphs above, there is no relationship between the players who played in at most 20 games and the 20 players who contributed the most to the score. We can check whether this statement is true or not and whether there is a relationship between them using correlation.

metin, ekran görüntüsü, renklilik, kare içeren bir resim

Açıklama otomatik olarak oluşturuldu

As can be seen from the correlation above, there is a positive and strong correlation of 0.78 between PTS and GP. Therefore, it cannot be said that there is no relationship between these two columns for the charts where we only look at the first 20.

**Data Scaling**

As we see the above “missingno” graph there is no null value. So we can check data values for scaling. There is looking so different between to values. for example, MIN has a 3028 and FG3\_PCT has a 0.351 value.

diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Even though the FG\_PCT column is normal data, this and similar columns are percentage columns, so this situation is expected. The other columns appear to be skewed data, so Min-Max Scaling should be applied for this.

**Feature Engineering**

At this stage, just because the value of our PLAYER column is object, LabelEncoder has been applied to the PLAYER column.

**Modelings**

The project includes regression models based on player statistics. Performance predictions were made using linear regression, Ridge regression and Losso regression.

For the purpose of our model, we separate the PTS column, that is, the total number of points scored by the players. Our aim here is to measure the impact of the players on the total score.

Data training and testing sets are divided into 80% and 20%.

Note: By comparing the players' scoring average and the overall scoring average, situations where players contribute more than the overall average are examined.

Additionally, according to the data set, the average score of the players is 368.

The graph of Total Number of Players and Number of Highly Contributing Players is as follows;

metin, ekran görüntüsü, dikdörtgen, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

According to this situation, the graphics and score information of the models we applied are as follows.

metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, çizgi, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Mean Squared Error (MSE) is a metric that measures how far a regression model's predictions are from the true values. MSE squares each forecast error, averages those errors, and so on to obtain a number. Lower MSE values ​​indicate better performance of the model.

**Conclusion**

When the graphic values and scores above are examined, it appears that linear and ridge regressions are overfitting. To avoid this situation, we should choose the lasso regression that best suits our data set.

**Project GitHub Link:** <https://github.com/SertacBattal/StratThink_Sept23_CA2>

**References**

Deloitte (2022). *Deloitte Football Money League | Deloitte UK*. [online] Deloitte United Kingdom. Available at: https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/deloitte-football-money-league.html [Accessed 26 Oct. 2023].

Forbes (2020). *The Business of Basketball: Forbes Releases 22nd Annual NBA Team Valuations*. [online] Forbes. Available at: https://www.forbes.com/sites/forbespr/2020/02/11/the-business-of-basketball-forbes-releases-22nd-annual-nba-team-valuations/#4ee34fa675ff [Accessed 26 Oct. 2023].

Gough, C. (2017). *NFL Revenue 2001-2017 | Statista*. [online] Statista. Available at: https://www.statista.com/statistics/193457/total-league-revenue-of-the-nfl-since-2005/ [Accessed 26 Oct. 2023].

Paulsen (2023). *NBA ratings deep-dive: Where does the league stand?* [online] Sports Media Watch. Available at: https://www.sportsmediawatch.com/2023/04/nba-ratings-viewership-past-30-years-analysis-where-league-stands/ [Accessed 26 Oct. 2023].

Statista Research Department (2023). *Total NBA league revenue 2001/02 to 2021/22*. [online] Statista. Available at: https://www.statista.com/statistics/193467/total-league-revenue-of-the-nba-since-2005/#:~:text=In%20the%202021%2F22%20season [Accessed 26 Oct. 2023].

Young, J. (2019). *Major League Baseball revenue for 2019 season hits a record $10.7 billion*. [online] CNBC. Available at: https://www.cnbc.com/2019/12/22/report-mlb-revenue-for-2019-season-a-record-10point7-billion.html [Accessed 26 Oct. 2023].

justplay (2023). Basketball Statistics Glossary. [online] Justplayss. Available at: https://justplayss.zendesk.com/hc/en-us/articles/115001859694-Basketball-Statistics-Glossary.

Schmitzhaus, V.M., Oliveira, W.G. and Almeida, M.B. de (2022). High-intensity Effort Impairs Basketball free-throw Shooting Efficiency. Motriz: Revista De Educação Física, 28. doi:https://doi.org/10.1590/s1980-657420220000422.

squared2020 (2017). Breaking Down Player Efficiency Rating. [online] Squared Statistics: Understanding Basketball Analytics. Available at: https://squared2020.com/2017/09/01/breaking-down-player-efficiency-rating/.

wolframcloud (2023). Basketball True Shooting Percentage | Wolfram Formula Repository. [online] resources.wolframcloud.com. Available at: https://resources.wolframcloud.com/FormulaRepository/resources/Basketball-True-Shooting-Percentage#:~:text=The%20true%20shooting%20percentage%20is.