Using Machine Learning Techniques to Predict Factors of Winning in the NBA

Author: Anthony Chan (500766830)

CIND820 DJ0- Big Data Analytics Project

Supervisor’s Name: Sedef Akinli Kocak

February 13, 2022

Table of Contents

[Abstract 3](#_Toc452457337)

[Literature Review 4](#_Toc452457338)

[Methodology 7](#_Toc452457340)

[References 12](#_Toc452457340)

# Abstract

Data analytics have impacted the sporting industry greatly since we can now analyze and predict what a player’s potential can be. Many of these statistics and factors can be translated to what we can see developing in the game. A game where each team can optimize their strategies to win. In this capstone project, the theme will be revolting around predictive analysis and knowledge discovery about the evolution of shooting the basketball within the NBA (National Basketball Association). The dataset used for this project is composed of statistics of NBA players and teams from the years, 1946 to 2016. However, I will be focusing on the players within the recent years of the dataset and collecting and comparing the data for these years. Many factors can go into winning a championship in the NBA, for example the 3-Point Field Goal Percentage, Field Goal Percentage, Free Throw Percentage, Turnovers, Offensive and Defensive Rebounds, and many more statistics.

The model that I will be creating will use the data provided to simulate the how the players in the NBA are shooting in the modern era and compare them to an older generation of basketball. This will allow us to find out the best performing model and use it as a basis for our study. Next, the model will attempt to predict how players will adapt to this new playstyle based on trends found which will help create insight to what teams and coaches should prioritize. This project would be able to investigate the problems: How important are each variable within the sport? Can we determine which factors affect winning chances of the team? Can we correlate a certain variable and relate it to a team’s chances of winning in the NBA? For this project, I will be using classification techniques and predictive modelling methods such as Logistic Regression, Naïve Bayes, and XGBoost. I will be implementing my analysis with software in Python where I will use packages such as Pandas, Numpy, SKlearn, Matplotlib, and others to compute predictive modelling techniques and visualize our data.

In conclusion, this capstone project will dive into the variables that take place in every NBA basketball game and learn about how they impact the players around the organization.

## Literature Review

Teams that use data analytics to track and predict data gives them a huge advantage to their competitors. This literature review will discuss what data analyst have studied within the NBA and how teams have taken advantage of these circumstances to determine a winning strategy.

The NBA has been an organization since 1946 and has been growing and developing to be one of the biggest sports in the world. There have been many changes within the rules and how the game has been played. For example, the league did not introduce the three-point shot until 1979 and how the league has changed from Centers playing near the basket to Guards shooting far from the basket. The game of basketball will forever be evolving as players grow up watching their favourite basketball player and they develop their own skills.

Many teams have started using data analytics to determine the outcome of their players. An example for this would be in an article by Wharton (2017) regarding the commissioner of the league. As mentioned in Wharton (2017), teams have started to monitor their players during practice to measure fatigue and prepare their players for any resting time. This will allow players to continue to stay healthy within the season and avoid injuries which should result teams to have a better win percentage. Another interesting fact from this article is that teams are using analytics to plan strategies. An exampled explained in Wharton (2017), 42% of the time a player goes forward with his left foot. This can lead to coaches developing strategies to counter this movement from players. Teams have been using data acquired from prospects in high school and college to determine how the future of the franchise by drafting players with certain aspects and skillsets. These are some of the external factors that can determine how teams perform during the season.  
 This capstone project will use statistics from players currently in the league. In the study shown from Lieder (2018), they investigated certain factors related to matches to create a model for predicting the results of NBA games. The authors used machine learning techniques such as logistics regression and linear regression to build predictive models to evaluate the outcome of these games. The model resulted in a 70% accuracy rating which the author has noted that this is extremely high as there can be multiple variables and aspects to the game that cannot be predicted.

Miljkovic (2010) studied machine learning methodologies such as k-nearest neighbor, decision trees, linear regression, and Naive Bayes where they would attempt to predict the outcome of NBA games. The dataset that they used were games from the years 2009 – 2010 and some examples of the variables used were free throws, three points, field goals, blocks, home games and away games, fouls, win percentage, and loss percentage. Their results using these machine learning models had an impressive 67% predictive accuracy. Like the study done in Lieder (2018), many variables that cannot be quantified affected the accuracy of the models.

Mikołajec (2013) conducted a study where they wanted to determine what teams in the NBA were doing that lead them to succeed. In the study, they worked with 52 variables such as Win percentage, Offensive efficiency, 3rd Quarter Points Per Game, Average Steals and Fouls. They determined that they would use the winner of the current season as their dependent variable and used the methods Pearson’s coefficient and regression analysis with the correlation matrix to identify optimal combinations of variables. Mikołajec (2013) noted that studies done in other basketball leagues such as the European Basketball League showed that the factors 3 points attempts, field goals percentage, free throws made, free throw percentage, defensive rebounds, and turnovers impacted a team’s chances of winning the most.

Another study done, Oliver (2005) had the factors field goal percentage, turnovers, offensive rebounds, and free throws made affected the results of the game. The reason for these 4 statistics to be so impactful is how they lead to advantages to the opposing team. For example, turnovers gave the opposing team more possessions and chances to score a basket. Rebounds allow the offensive team to have a second chance at scoring. Additionally, causing a significant number of free throws to a team can affect how the coach makes substitutions and plays.

As shown from multiple articles, this type of study has been done before as analytics in the sporting industry is growing exponentially. However, this literature review will also highlight similarities and differences between this project and others. I will be focusing on different research questions as I would like to take on the importance of certain variables and I will also be using different seasons of the NBA to differentiate my study against others. Based on the articles I have reviewed; I will be using XGBoost to assist with classification methods as it is a decision tree based machine learning algorithm. Logistic regression and Naïve Bayes have been a standard model used in majority of the studies shown from the articles. For that reason, I will be using those predictive modelling methods to understand the relationship between variables and be able to compute predictions. Since this study will be based on the same material, I will be able to this capstone project with studies that have been done in the past to further prove or disprove my research questions.

**Data Description**

The dataset used consists of 21476 rows of players in the NBA which consists of 29 variables. However, since many rules were not implemented in the early years of the NBA. I have removed years from 1947 – 1992 and we have a remaining of 10706 rows of players consisting of the same variables. In Table 1, a brief description of the variables used is shown.

Table 1 – Variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variables | Description | Datatype | Type of Variable | Mean | Median |
| Player | Name of Player | Object | Categorical |  |  |
| Season | Current Season Data | Object | Categorical |  |  |
| Age | Age of Player | Float64 | Numerical | 26.951429 | 26.000 |
| Tm | Team | Object | Categorical |  |  |
| Lg | League | Object | Categorical |  |  |
| G | Games Played | Int64 | Numerical | 53.495984 | 61.000 |
| GS | Games Started | Float64 | Numerical | 26.225481 | 11.000 |
| MP | Minutes Played | Float64 | Numerical | 1268.281431 | 1178.000 |
| PER | Player Efficiency Rating | Float64 | Numerical | 12.749491 | 12.900 |
| 3PAr | 3-Point Field Goal Percentage | Float64 | Numerical | 0.191115 | 0.143 |
| FTr | Free Throw Rate | Float64 | Numerical | 0.315403 | 0.284 |
| ORB% | Offensive Rebound Percentage | Float64 | Numerical | 6.074007 | 5.000 |
| DRB% | Defensive Rebound Percentage | Float64 | Numerical | 14.110221 | 13.200 |
| TRB% | Total Rebound Percentage | Float64 | Numerical | 10.093880 | 9.200 |
| AST% | Assist Percentage | Float64 | Numerical | 12.987041 | 10.100 |
| STL% | Steal Percentage | Float64 | Numerical | 1.644969 | 1.500 |
| BLK% | Block Percentage | Float64 | Numerical | 1.565290 | 1.000 |
| TOV% | Turnover Percentage | Float64 | Numerical | 14.553889 | 13.700 |
| USG% | Usage Percentage | Float64 | Numerical | 18.841316 | 18.500 |
| ORtg | Offensive Rating | Float64 | Numerical | 101.932092 | 104.000 |
| DRtg | Defensive Rating | Float64 | Numerical | 106.110623 | 106.000 |
| OWS | Offensive Win Shares | Float64 | Numerical | 1.398608 | 0.600 |
| DWS | Defensive Win Shares | Float64 | Numerical | 1.295068 | 0.900 |
| WS | Winshares | Float64 | Numerical | 2.694648 | 1.700 |
| WS/48 | Winshares per 48 minutes | Float64 | Numerical | 0.072558 | 0.081 |
| OBPM | Offensive Box plus minus | Float64 | Numerical | -1.572492 | -1.300 |
| DBPM | Defensive Box plus minus | Float64 | Numerical | -0.431048 | -0.400 |
| BPM | Box plus minus | Float64 | Numerical | -2.003307 | -1.600 |
| VORP | Value Over Replacement Player | Float64 | Numerical | 0.640379 | 0.100 |

Text

Description automatically generated with low confidence

Table 2 – Correlation between variables

In Table 2 is a quick correlation between each the variables.

Application, table, Excel

Description automatically generated

**Methodology**

Table 3 shows a flow chart of the methodology that will be done throughout this capstone project.

A picture containing text, screenshot, businesscard

Description automatically generatedTable 3

#### References

Khan, E. (2017, October 3). *Advanced NBA stats for dummies: How to understand the new hoops math*. Bleacher Report. Retrieved February 2, 2022, from https://bleacherreport.com/articles/1813902-advanced-nba-stats-for-dummies-how-to-understand-the-new-hoops-math

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. Annals of Data Science, 6(1), 103-116.

Wharton (2017, June 1). *The NBA's Adam Silver: How Analytics is transforming basketball*. Knowledge@Wharton. Retrieved February 8, 2022, from https://knowledge.wharton.upenn.edu/article/nbas-adam-silver-analytics-transforming-basketball/

Grant , H. (2015, July 23). How the game has changed. Retrieved February 8, 2022, from https://www.horacegrant.com/blog/how-the-game-has-changed/

Lieder, N. (2018, July 25). *Can machine-learning methods predict the outcome of an NBA game?* SSRN. Retrieved February 9, 2022, from https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3208101

Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010, September). The use of data mining for basketball matches outcomes prediction. In IEEE 8th international symposium on intelligent systems and informatics (pp. 309-312). IEEE.

Mikołajec, K., Maszczyk, A., & Zając, T. (2013). Game indicators determining sports performance in the NBA. Journal of human kinetics, 37, 145.

Csátaljay, Gábor & James, Nic & Hughes, Mike & Dancs, Henriette. (2013). Effects of defensive pressure on basketball shooting performance. International Journal of Performance Analysis in Sport. 13. 10.1080/24748668.2013.11868673.

Oliver, D. (2005). What wins basketball games, a review of „Basketball on paper: Rules and tools for performance analysis”. Polomac Books, 26-85.

Wen, R. (2017, September 17). NBA Data. figshare. Retrieved January 23, 2022, from

<https://figshare.com/articles/dataset/NBA_data/5414170>

Github - <https://github.com/anthonyychan/cind820>