What Nba Players Statistics Best Predict Scoring Output.*

A Look Into The Metrics Which Best Predit Scoring Output Among NBA Players From The 2020-21 Season.

Adam Labas

25 April 2022

Abstract

First sentence. Second sentence. Third sentence. Fourth sentence.
Keywords:NBA, PTS, REB, AST, MIN, FGA, X3PM

Contents

*Code and data are available at: https://github.com/adam-labas/Which-NBA-Stats-Best-Predict-Scoring-Output.

1 Introduction

The NBA has for long been a widely admired and celebrated facet of American popular culture. Every season, the average fan has the opportunity to watch the best players on the face of the earth go at each other, night after night, in nail-biting intensity. As an avid fan of the NBA, I too am a consumer of the NBA and its professional basketball content. This became foundational in my life in 2019 when my hometown Toronto Raptors had an amazing run in the playoffs and won the NBA championship. The NBA is well known for being a league full of tall players and as my data suggests, this is true. However, the Toronto Raptors team that won the NBA finals had two relatively short players: Fred VanVleet and Kyle Lowry. Being a relatively short man myself, I was mesmerized by their abilities to perform at the highest level and their abilities to score the basketball and impact the game. As such, in this Paper, I attempt to explore the true parameters which contribute to players scoring output.

As the new 2021-2022 NBA season started on October 19th 2021, it was reminiscing and was thinking about traditionally short players and their ability to score the basketball with ease. It is counter intuitive to me to think that a player that is 185 cm and 89kg like VanVleet can score with ease on a player that is much taller and heavier than him.

If a player like VanVleet, lacking in height can score the ball with ease, what are actually the qualities and traits which contribute statistically to players being able to score more points per game? In statistical terms what predictors best describe the points per game of an NBA player in the 2020-21 season? We aim to answer this question by developing a simple, yet effective and easy to understand model.

2 Data

In this Data Section 2, I will provide a look into the data acquisition and processing methodology as well as a deep dive into the contents of the data. We will also touch on our exploratory data analysis as it pertains to variable selection and lastly we will discuss the reach of the data.

First of, (Table 1) give us a glimpse of the data.

Table 1: First ten rows of a dataset of shelture usage without All Population

Player	Points	Rebounds	Age	Minutes	Field Goal Attempts	3PTS Made	Weight	Height
Aaron Gordon	12.4	5.7	25	27.7	10.0	1.2	107	203
Aaron Holiday	7.2	1.3	24	17.8	6.6	1.0	84	183
Aaron Nesmith	4.7	2.8	21	14.5	3.9	0.9	98	196
Abdel Nader	6.7	2.6	27	14.8	4.8	0.8	102	196
Adam Mokoka	1.1	0.4	22	4.0	1.4	0.1	86	193
Al Horford	14.2	6.7	35	27.9	12.9	2.0	109	206
Al-Farouq Aminu	4.4	4.8	30	18.9	4.3	0.3	97	198
Alec Burks	12.7	4.6	29	25.6	10.2	2.1	86	213
Aleksej Pokusevski	8.2	4.7	19	24.2	9.1	1.3	98	208
Alen Smailagic	1.9	1.1	20	5.6	1.8	0.3	84	193

2.1 Data Collection

All the data used in this analysis was retrieved directly from the NBA's website and used in accordance with their terms and conditions (more on this in section 5.4). The data on the NBA's website is displayed on 11 different pages each containing 50 players except for the last page which only contains 36 players. The data is available online as an HTML table. As a result, I used a Google Chrome web browser extension

titled **Dowload Table as CSV** to extract the data into 11 CSV files. More information on the Google Chrome extension and the source code that makes it work can be found at the following git repository: https://github.com/arktiv/table-csv-chrome. After appending all the player data into one dataframe, we now have a dataset

Our data is of penguins (Figure **??**).

2.2 EDA: Exploratory Data Analysis

As was stated in the Introduction Section 1, the aim of this analysis is to find a parsimonious model

As a result, I had originally decided to study the relationship between the Age, Height and Weight and the number of points scored. When plotting my data, I noticed that there was not a linear relationship between the predictor variables I chose to study and the response variable. As a result, I modified my research question to instead study the relationship between points per game (PTS) to minutes per game and assists per game as this will indirectly answer the question I had originally intended to answer. I will be able to make this link once I find data on the positions of players and demonstrate that players with smaller heights are almost always players with positions that traditionally have a lot of assists like the Point Guard.

I think that any analysis which gives an insight on which players are prone to producing the most points is always useful and an have impacts in many fields such as the world of sports gambling and especially fantasy sports. Although I am not a gambler myself, I am an aspiring actuary and data analyst and I find pleasure in being able to bring forth simple results from large complex datasets.

2.3 Population, Frame or Sample

It is important to discuss the reach of our data as this will have an effect on the generalizability of our Model. As previously mentioned, the City of Toronto funds services for people experience homelessness and the data we have acquired for this paper reflects all the data collected in the SMIS system in the years 2020 and 2021. This is important because it brings to light one of the first limitations of the data which we will discuss more in Section **??**. Although the SMIS data is quite expansive with 323 data points, it fails to capture the entire population of homeless people in the city of Toronto as it does not take into consideration homeless people who sleep on the street or those only using services which are not government funded. Thus, in this paper, we are only deal with a sample of the population, but a big one nonetheless.

3 Model

$$Pr(\theta|y) = \frac{Pr(y|\theta)Pr(\theta)}{Pr(y)} \tag{1}$$

Equation (1) seems useful, eh?

We can use maths by including latex between dollar signs, for instance θ.

4 Results

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Limitations And Weaknesses

5.5 Next Steps

itll be intersting to see how our model changes potentially if we only include players that meet a certiann threshhold. Like maybe only players who have playes x many games or shoot at least 5 times.

Appendix

A Additional details

B References