The Secret Formula to Winning Basketball

By: Gurjus Singh

12th March 2020

MSDS 430: Introduction to Data Science

Introduction

Basketball has been one of my passions since I was eleven years old. I started watching and learning about the Sacramento Kings at age eleven. I wanted to imitate the players' moves, and had a passion and love for them that did not go away. As I grew older I wanted to play in the NBA and specifically with the Kings, but unfortunately did not have the height to actually make it as I stood at 5'3". As I started my senior year of high school in 2013 and college in 2014, I wanted to do something that would allow me to work in sports. At this time the NBA, started to go through this phase of Sports Analytics with Vivek Ranadive taking over as owner of the Sacramento Kings (Xu, 2014). His plan was to use technology to move the Kings into the future (Xu, 2014). Other owners such as former CEO of Microsoft, Steve Ballmer, also bought an NBA team in 2015, the L.A. Clippers moving the NBA closer to the age of Analytics (Forbes, n.d.). I also researched Daryl Morey at this time, who was once a Northwestern student like myself, who used Analytics to build the Houston Rockets that are now competing with the Golden State Warriors for a NBA Championship (Ubillis, 2018; ESPN, 2017).

All these people have motivated me to become a data scientist to allow me to not just work in any industry that requires data, but also in sports. With my data science knowledge, I can help an NBA team find the next great NBA player. I have current NBA experience with analyzing different player's stats, but do not have the necessary data science skills to better understand these stats. With this project I am hoping to understand how the tools can be useful, and what attributes I can take into account that I have in the data.

The data file came from Kaggle.com and includes data from the NBA seasons from the years 2016-2019. The data in the file was specifically extracted from sites such as Basketball Reference, a home for statistics from every player in every NBA game, Hoops Hype, where

salary data and breaking news is available, and another Kaggle data file which contained Wikipedia views from each player's page. This data file was specifically chosen, as it has relevant statistical categories to find the best players in the NBA and what makes a winning player. I am hoping to find statistical categories that can be associated with winning players and teams.

Data preparation and analysis

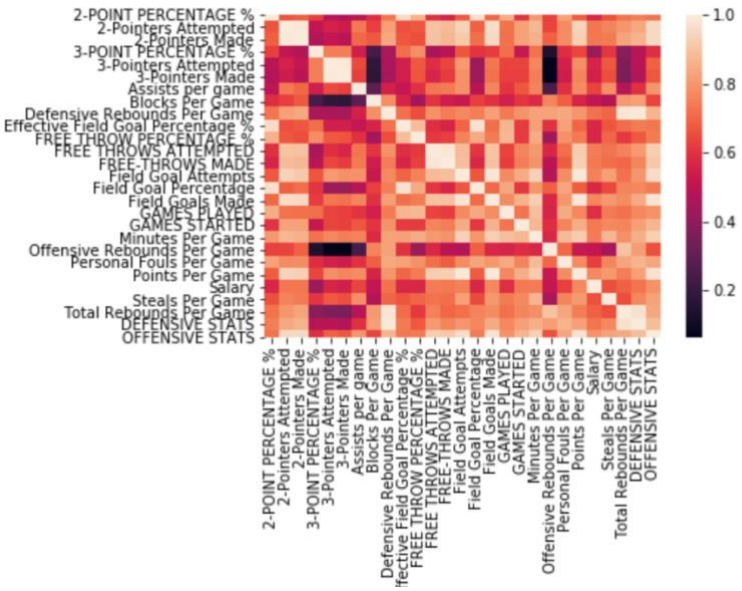
Before doing my initial analysis, I decided to clean up the data to take out the less salient attributes in the data. For my original it included all of the attributes below:

- Rk - type- int, describes how the player is in comparison to the rest of the NBA.
- Player_ID - unique ID provided for each player, type- string
- Conference - which conference a player is from, type -string
- Player.x - shows the player's full name in type- string
- Pos1 - player's primary playing position, type- string
- Pos2 - player's secondary playing position, type- string
- Age - player's age, type- int
- Tm - player's team abreviation of city, type- string
- G - Games played in for each player, type- int
- GS - Games started in for each player, type- int
- MP - Minutes played per game, type- float
- Role - signifies what lineup they play in such as are they a tall player, which means they play in Front Court, or do they do scoring, and are they short, so they are Back Court, type- string
- Fvot - signifies how many votes they received by fans for all star game - type int.
- FRank - ranking in terms of fan vote, type - float.
- Prank - player vote ranks for all star game
- Pvot - player vote by team captains for all star game
- Mvot - coach votes received for all star game
- Mrank - coaching vote rank received for all star game
- Score - score received in terms of all star votes
- Play - signifies whether player played in all star game, YES OR NO? type- string
- FG - field goals made per game, type float
- FG. - filed goal percentage, type float
- FGA - field goals attempted made per game, type float
- X3P - 3 pointers made per game, type float
- X3PA - 3 pointers attempted per game, type float
- X3P. - 3 point percentage, type float
- X2P. - 2 point percentage, type float

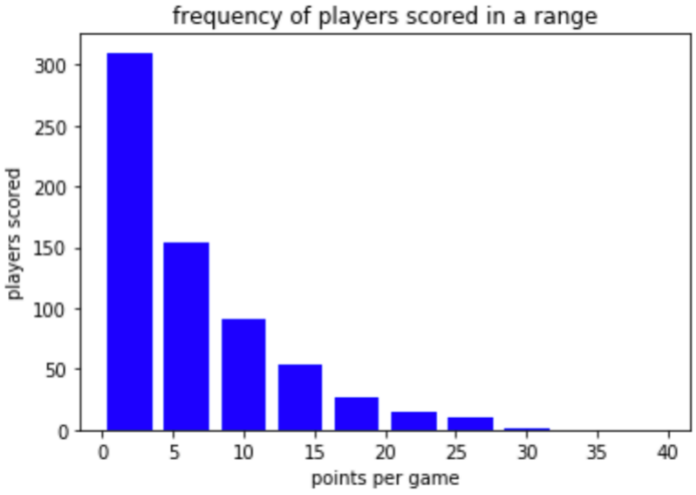
- X2PA - 2 point attempts per game, type float
- X2P - 2 point makes per game, type float
- eFG - effective field goal percentage, type float
- FTA - attempts per game, type float
- FT - free throws made on average, type float
- ORB - signifies offensive rebounds a player got, type float
- DRB - defensive rebounds a player grabs, type float
- TRB - total rebounds a player grabs, type float
- AST - assists a player makes, type float
- STL - steals per game, type float
- BLK - blocks a player makes defensively, type float
- TOV - turnovers a player commits, type float
- PF - personal fouls a player receives, type float
- PTS - points per game, type float
- Salary - money a player makes, in millions, type float
- mean_views - views each player receives off their wikipedia page, type int
- Season - what season are the averages associated in, type string

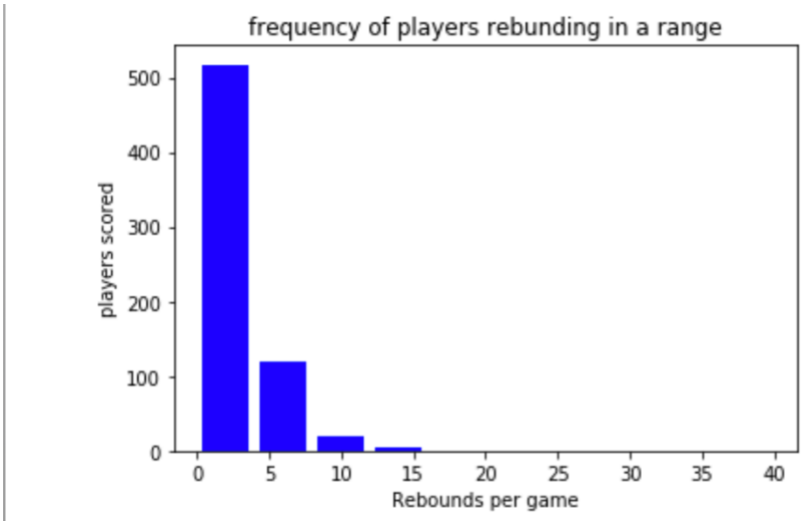
Specific attributes that I decided to drop from the table were Season, mean_views, Rk, Conference, Fvot, Frank, Prank, Pvot, Mvot, Mrank, and Team. I chose these as less salient features because they did not affect how a player contributed to winning. After I dropped these less salient features, I decided to combine the season averages in all statistical categories over all 3 seasons into one average for every statistical category that way it was easier to proceed with the analysis.

For my first visual, I decided to use a Heat Map to see if there were any surprising correlations. I did not find anything surprising, but rather found correlations that made sense such as between points per game and minutes per game. The positive correlation meant that there was a relationship, usually linear between minutes and points. This was true because as more minutes are played the player usually scores more points.



I proceeded to move on to some bar graphs. I found some interesting findings in that less that 100 players actually averaged 20+ point over 3 seasons. I also noticed that less 50 players averaged 10+ rebounds again. I decided to show my findings in these statistical categories because these are statistics that many winning teams tend to do well in. I wanted to see which players had both of these types of statistics. I decided the best way to find out was by making new variables.



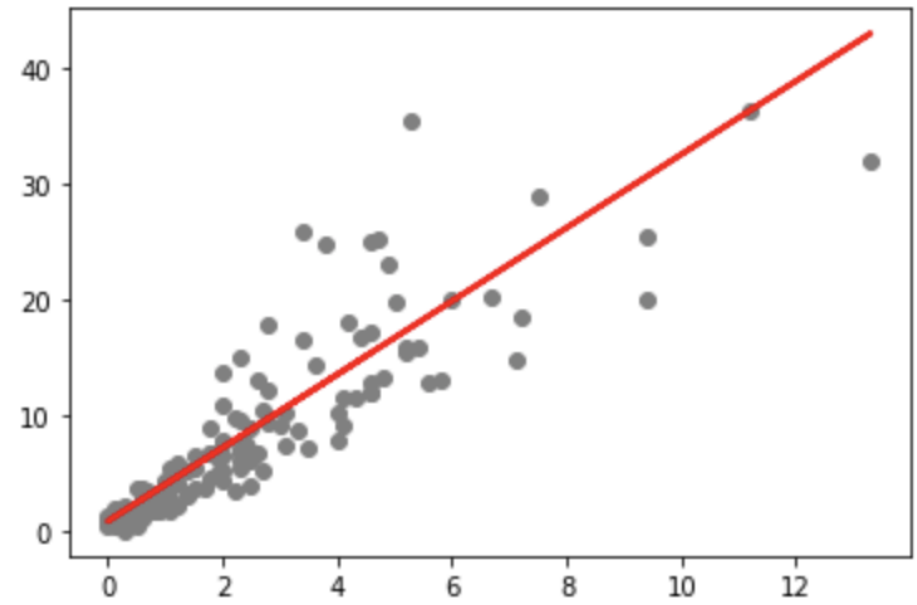


Two variables I decided to make in my analysis were Offensive Stats, and Defensive Stats. As the name implies, Offensive Stats were stats combined from offensive categories such as points per game, assists per game, and offensive rebounds, while defensive stats were defensive rebounds, steals, blocks. I then decided to combine these two new variables into one variable by adding the combined stats together and found that the variables in fact were great at finding the top players.

	Name	OffandDef
358	Russell Westbrook	53.3
174	James Harden	52.0
24	Anthony Davis	50.9
145	Giannis Antetokounmpo	48.8
260	LeBron James	48.7
...	...	...
546	Reggie Hearn	0.3
658	Zach Lofton	0.3
478	Erik McCree	0.2
488	Jacob Pullen	0.2
591	Donte Grantham	0.0

660 rows × 2 columns

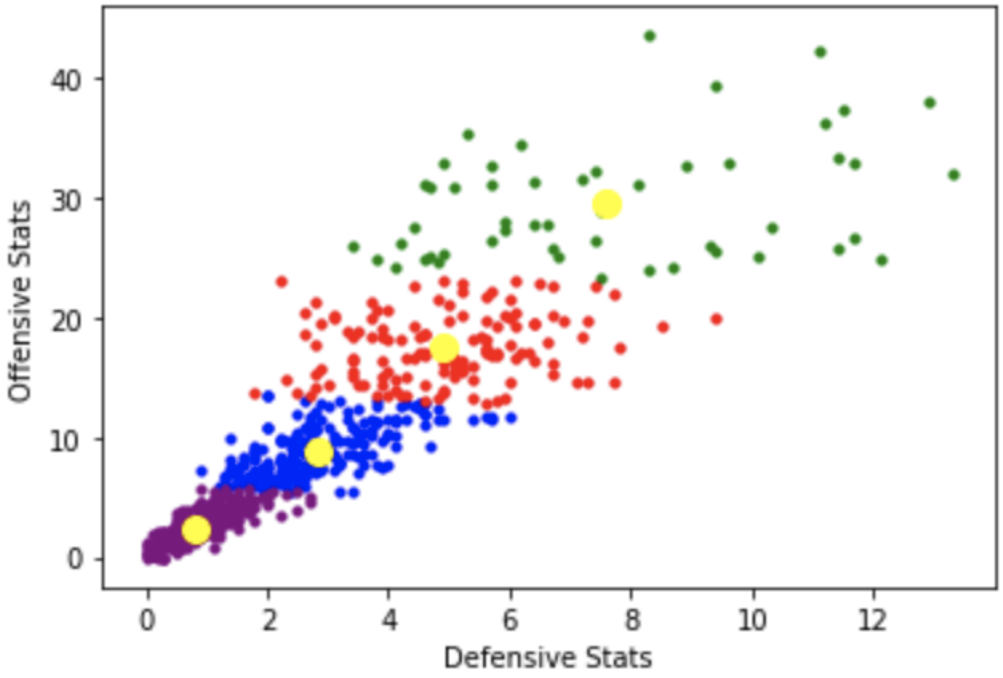
The variables combined quickly and accurately identified the top five great players in the league, all of which are on the top 3 teams in the NBA. I further did more analysis, by using linear regression and clustering.



In my regression algorithm, it accurately predicted what future NBA players needed to do to be successful. Specifically, it was saying, that future NBA player need to show they can play both defense and offense. My two new created variables had a positive correlation of 0.791. This was surprising, but the more I thought about it the more it made sense, because defense is a way to score more easily as defense can create turnovers, and turnovers can create easy opportunities to score.

I lastly wanted to find out if a clustering algorithm could categorize the top players. In the clustering algorithm it categorized the top players as green in the visualization below. I wanted to find out specifically which players were in the green. It categorized a total of 51 players as top players. As I read the list, I was surprised to see names like Dwight Howard and Enes Kanter who at that time were not considered top players and are also not today. This shows although I

think that my clustering algorithm is correct, I did not account for minutes per game as this should take into account which players are efficient based on minutes.



17	Andre Drummond
23	Andrew Wiggins
24	Anthony Davis
35	Blake Griffin
41	Bradley Beal
66	Chris Paul
68	CJ McCollum
69	Clint Capela
76	D'Angelo Russell
79	Damian Lillard
95	DeAndre Jordan
99	DeMar DeRozan
100	DeMarcus Cousins
103	Dennis Schröder
110	Devin Booker
123	Dwight Howard
130	Enes Kanter
131	Eric Bledsoe
145	Giannis Antetokounmpo
147	Goran Dragic
152	Hassan Whiteside
174	James Harden
195	Jimmy Butler
203	Joel Embiid
208	John Wall
221	Jrue Holiday
223	Julius Randle
228	Jusuf Nurkic
230	Karl-Anthony Towns
231	Kawhi Leonard
234	Kemba Walker
238	Kevin Durant
239	Kevin Love
243	Klay Thompson
250	Kyle Lowry
254	Kyrie Irving
255	LaMarcus Aldridge
260	LeBron James
261	Lou Williams
272	Marc Gasol
296	Mike Conley
314	Nikola Jokic
316	Nikola Vucevic
330	Paul George
331	Paul Millsap
357	Rudy Gobert
358	Russell Westbrook
379	Stephen Curry
382	Steven Adams
422	Victor Oladipo
433	Zach LaVine

Conclusion

Although, some players in my analysis I did not consider elite players, most of them are considered high impactful players on their teams. Although this does show players that are doing well statistically, this sometimes does not tell the whole story. Players that are doing well, sometimes cannot make their teams win alone. It takes a combination of players, to make a team win.

For example, the Warriors during their championship run relied on players such as Klay Thompson, Stephen Curry and Kevin Durant (Reference, n.d.). Also, sometimes statistics do not provide the whole story to a team's success. Other factors such as player happiness, team chemistry which is referred to as player camaraderie, injuries, team budget, and ownership can have effects on winning. It is important to incorporate these using predictive models. In my case I was not provided with sufficient data to take into account injuries. Also, most of these players have been in the NBA, since eighteen to twenty years old, so my analysis did not take this into account. Regular NBA general managers have a challenge of finding basketball players from college and high school. Even though some basketball players in college or high school do have a successful career and some do not, there is still uncertainty to how successful they can be in the NBA, as majority of the players are muscularly and mentally stronger in the NBA. This shows finding the best players is not an "exact science" as originally anticipated (Adamek, 2017, para 8).

In my future analysis of NBA players and this data set, I look forward to accounting for how minutes can affect each players stat. I also want to dedicate time to build a fantasy team, using stats and constraints to a budget. With this budget constraint I will be able to identify which players are being overpaid and which ones are being underpaid. I can simulate how this

team I made will perform through the video game NBA 2K which was made to replicate the environment of a real NBA game (Wikipedia, 2020).

References

Adamek, S. (2007, June 26). NBA draft not an exact science, so check back in five years. Retrieved February 29, 2020, from https://journalstar.com/sports/nba-draft-not-an-exact-science-so-check-back-in/article_e61fd9e5-c0b2-5ee8-80ea-115e07f89ce0.html

Kaggle. (2019, December). NBA players 2016-2019. Retrieved February 28, 2020, from https://www.kaggle.com/davra98/nba-players-20162019

ESPN. (2017, December 22). Daryl Morey: beating Golden State 'the only thing we think about'. Retrieved from https://www.espn.com/nba/story/_/id/21841648/daryl-morey-houston-rockets-says-team-obsessed-beating-golden-state-warriors

Forbes. (n.d.). Steve Ballmer. Retrieved February 28, 2020, from https://www.forbes.com/profile/steve-ballmer/#7ecb74c14818

Kaggle. (2019, December). NBA players 2016-2019. Retrieved February 28, 2020, from https://www.kaggle.com/davra98/nba-players-20162019

Reference, B. (n.d.). 2017-18 Golden State Warriors roster and stats. Retrieved February 28, 2020, from https://www.basketball-reference.com/teams/GSW/2018.html

Ubillus, C. (2018, January 18). Houston Rockets general manager Daryl Morey talks basketball, data analytics at CTSS event. Retrieved February 28, 2020, from https://dailynorthwestern.com/2018/01/17/campus/houston-rockets-general-manager-talks-basketball-data-analytics/

Wikipedia. (2020, February 25). NBA 2K. Retrieved February 28, 2020, from https://en.wikipedia.org/wiki/NBA_2K

Xu, A. (2014, February 26). Vivek Ranadive gives Sacramento Kings new direction with technology. Retrieved February 28, 2020, from https://www.sporttechie.com/vivek-ranadive-gives-sacramento-kings-new-direction-with-technology/?theme=active