JANUARY 30, 2023

A DEEP DIVE INTO THE NATIONAL BASKETBALL ASSOCIATION (NBA) THROUGH THE LENS OF DATA SCIENCE

THREE-POINT SHOOTING TRENDS, PLAYER POSITIONINGS, AND MVP PREDICTIONS

TONY (FENG-CHI), CHU
UNIVERSITY OF BATH
Msc in Business Analytics

I. Introduction

The National Basketball Association (NBA) has a long history dating back to 1891 when basketball was invented by Dr. James Naismith, a physical educator and sports coach at Springfield College. It wasn't until 1946, when Boston Garden owner Walter Brown recognized the potential for hosting basketball games in ice hockey arenas during the off nights, that the NBA was founded. Brown established the Basketball Association of America (BAA), which merged with the National Basketball League (NBL) in 1949 to become the NBA. In 1976, the NBA merged with the American Basketball Association (ABA), adding four franchises to the league. This solidified the foundation of the league which has gained global popularity over the years.

The NBA experienced a difficult start as it struggled with financial problems and a lack of popularity. However, during the civil rights movement in the 1960s, media attention on the mistreatment of black NBA players, including iconic figure Bill Russell, helped increase the league's national popularity. The athletic prowess of Wilt Chamberlain, the rivalry between Larry Bird and Magic Johnson, the dominance of Michael Jordan, the sustained excellence of LeBron James, and the current three-point revolution led by Stephen Curry have all contributed to the continued success and growth of the National Basketball Association. Internationally, the Olympics, the two World Wars where American troops helped spread the basketball culture, the injections of basketball talents all around the world, and the impact of social media all attributed to the NBA’s growing viewership. To reflect on the state of NBA basketball games, on top of how far it has progressed and evolved, it is necessary to acknowledge the explosive growth of sports analytics that helped foster a transcending approach to basketball games.

Though may come as a surprise to some people, Stephen Jay Gould, a late American paleontologist, and evolutionary biologist made a significant contribution to the field of sports analytics through his study on the decline of "the .400 hitter" in Major League Baseball (MLB). In baseball, a .400 batting average is considered extremely superior, and such performance surprisingly has not been achieved since the 1940s. By using statistical analysis and historical data, Gould was able to identify the factors that led to the decline of players with such high batting averages and demonstrate the value of using analytical approaches to understand and analyze sporting phenomena. His work inspired many sports teams to adopt similar approaches to improve their performances, including the Oakland Athletics’ astonishing winning streak in the early 2000s under the leadership of general manager Billy Beane, the New England Patriot’s impeccable player evaluation and drafting strategies, and the Gold State Warriors innovative use of analytics that helped build one of the most dominant NBA dynasties in the recent decades.

To contextualize the use of sports analytics and look at the evolution of NBA games from a statistical standpoint, this dissertation seeks to deep dive into different NBA data from various angles, including in-game performances, shot locations, all-star, and MVP players selections, and many more. The extensive research and analysis conducted in this dissertation are distilled into a few key points and components to address and answer the three main research questions below.

The first research question examines the NBA's three-point shot evolution. The NBA has undergone numerous changes and reached several important milestones, one of which was the introduction of the 3-point line in the 1979–1980 season, which was believed to increase scoring and make games more exciting. Without looking at the statistics, It seems that the NBA has changed over time, with teams now attempting more three-point shots per game to level the playing field. In the past, teams took fewer three-point shots per game. To explore the changes and the progressions of the game statistically from this angle, the dissertation solely looks at the in-game performances starting from the 1979–1980 season when the three-point line was inducted into the game so conclusions of the first research question may be drawn from the same baseline data.

The next research question focuses on player categorizations. Conventionally, there are five positions on the court, point guard, shooting guard, small forward, power forward, and center. Despite how each position performs on the court changes through time, the five positions mentioned technically have specific responsibilities and skills associated with each of them. The point guard is usually the primary ball handler and playmaker, and the one who set up the offense for each possession that leads to scoring opportunities. The shooting guard is someone who provides the scoring ability and impacts the game offensively. The great Michael Jordan is a prime example of a typical shooting guard. The small forward is the most versatile position on the team and can play both offense and defense, and it usually can play multiple positions due to its all-around skill sets. The power forward is technically someone who plays closer to the rim and is known for their rebounding and scoring abilities. Lastly, the center is usually the tallest player and is known as the defensive anchor of the team.

In modern basketball, players tend to have more diverse skill sets and can play multiple positions as the game has evolved. The league has also trended towards a more "positionless" style of play in recent years, due to the emergence of innovative franchise players who have redefined the game. By approaching the subject with fresh ideas and contrasting traditional roles with classifications based on statistical analysis, a new perspective on how positions in the NBA can be redefined in the current era can be achieved.

The final focus of the research will be on identifying the characteristics that make certain players stand out as being particularly talented or successful in the NBA. This involves examining patterns of individual player success, particularly on whether a player is more likely to be named the Most Valuable Player (MVP). The goal of this research is to understand whether factors contributing to a player's greatness can be identified and whether predictive models can be developed to foresee patterns that may be indicative of an MVP winner.

Given that the NBA collects an abundance of in-game data, this dissertation shall have enough data available for thorough analyses. But before jumping into the research methods and analyses themselves, it would be sound to review several relevant literature studies that cover the research questions mentioned above.

II. Literature Review

Early conversation of sports analytics

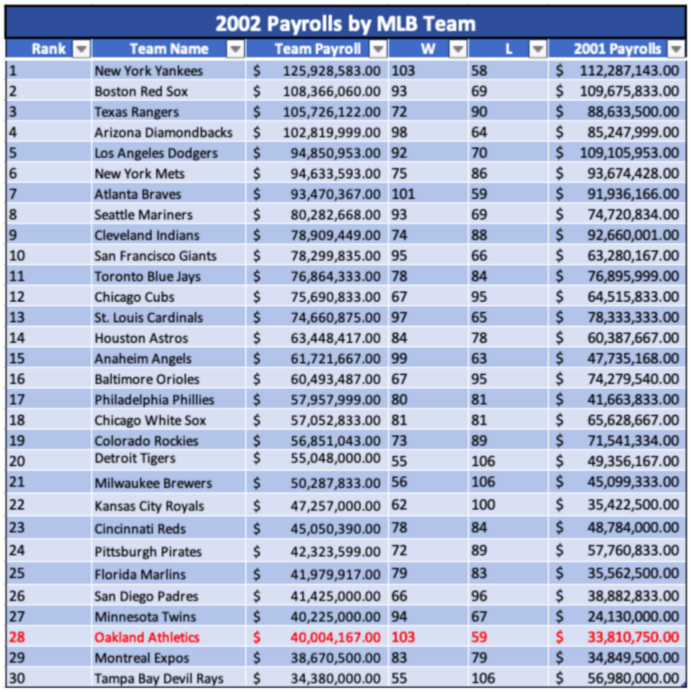
To provide more context and background information regarding the essence of this dissertation, it is crucial to deep dive into the origin of sports analytics and the impact sports analytics have on modern sports leagues. As introduced briefly in the introduction, Stephen Jay Gould, in his book published titled, Full House - The Spread of Excellence from Plato to Darwin, proposed that the evolution of sports does not follow the traditional model of natural selection, in which the fittest members of a population survive, and the unfit members are eliminated. Instead, he argued that sports evolve through random changes rather than by a directed and purposeful progress toward higher complexity (Gould, 1996). Gould suggested that there is an invisible right wall of biological limitations for humans in sports. Athletes approach these limitations steadily, and the performance gaps between players or variations decrease over time. As a zealous baseball fan, Gould was fascinated by the phenomenon that the world has yet to see a .400 hitter after Ted Williams from the 1941 Boston Red Sox achieved such a feat. Intuitively, it is easy to make assumptions or reach a conclusion without fully considering all the evidence or information that the performance of the MLB hitters is progressively declining. However, the findings suggest otherwise.

Some measures of batting performances throughout the decades are introduced in Table 1. From the table, it is evident that the batting average has remained constant, while the standard deviation continues to decline. Furthermore, the standardized scores suggest that the best performances of both the present and the past are as valuable. Ty Cobb's 0.42 batting average from 1911 was quite comparable to George Brett's 0.39 batting average from 1985 if we refer to the standardized score. It is observed that the batting performance remains constant, while the performance gaps between players continue to close as the overall standard of plays has continued to improve, as the findings showed in Chatterjee & Lehmann's article Evolution of team sports: a case study for National Basketball Association (1997). Inferences drawn from such findings suggested that it is important to recognize that because the overall level of play in the league has improved, it has become more challenging for players to stand out in their hitting performance. Moreover, collectively a system can expect improvements while attaining stability but with a moderate number of extremes. Such a conclusion would not have been drawn if we were to only consider the batting averages, and the implementation of analytics helps to settle the argument.

Decade	M	SD	Highest Average	Standardized Score
1910	0.266	0.3710	0.420 (Ty Cobb, 1911)	4.26
1940	0.267	0.0326	0.406 (Ted Williams, 1941)	4.42
1980	0.261	0.0317	0.390 (George Brett, 1985)	4.16
1990	0.270	0.0316	0.394 (Tony Gwynn, 1994)	3.47

Table 1. Batting performances of selected MLB seasons

The outcome of Oakland Athletics' 2002 season is another prime example of how impactful analytics play in professional sports leagues. Billy Beane, the Oakland Athletics general manager utilized advanced data analytics; Sabermetrics, to statistically analyze the data in baseball which aims to quantify MLB baseball players' performances based on objective statistical measurements, as Jacob Moorefield suggested in his Thesis, The Oakland Athletics use of sabermetrics and the rise of big data analytics in business (2021). An illustration drawn from the thesis as shown in Table 2 implied that a small market team with bottom-tiered payrolls like the Oakland A's was unable to attract tier-1 talents to play for them, resulting in the Oakland A's management thinking outside of the box. Instead of targeting productive individual players, Billie Beans and his player analyst Peter Brand, adapted by reconstructing the roster by targeting a combination of undervalued players using player analytics that would eventually give the team a competitive edge collectively.



2002 Payrolls by MLB Team

Rank	Team Name	Team Payroll	W	L	2001 Payrolls
1	New York Yankees	$ 125,928,583.00	103	58	$ 112,287,143.00
2	Boston Red Sox	$ 108,366,060.00	93	69	$ 109,675,833.00
3	Texas Rangers	$ 105,726,122.00	72	90	$ 88,633,500.00
4	Arizona Diamondbacks	$ 102,819,999.00	98	64	$ 85,247,999.00
5	Los Angeles Dodgers	$ 94,850,953.00	92	70	$ 109,105,953.00
6	New York Mets	$ 94,633,593.00	75	86	$ 93,674,428.00
7	Atlanta Braves	$ 93,470,367.00	101	59	$ 91,936,166.00
8	Seattle Mariners	$ 80,282,668.00	93	69	$ 74,720,834.00
9	Cleveland Indians	$ 78,909,449.00	74	88	$ 92,660,001.00
10	San Francisco Giants	$ 78,299,835.00	95	66	$ 63,280,167.00
11	Toronto Blue Jays	$ 76,864,333.00	78	84	$ 76,895,999.00
12	Chicago Cubs	$ 75,690,833.00	67	95	$ 64,515,833.00
13	St. Louis Cardinals	$ 74,660,875.00	97	65	$ 78,333,333.00
14	Houston Astros	$ 63,448,417.00	84	78	$ 60,387,667.00
15	Anaheim Angels	$ 61,721,667.00	99	63	$ 47,735,168.00
16	Baltimore Orioles	$ 60,493,487.00	67	95	$ 74,279,540.00
17	Philadelphia Phillies	$ 57,957,999.00	80	81	$ 41,663,833.00
18	Chicago White Sox	$ 57,052,833.00	81	81	$ 65,628,667.00
19	Colorado Rockies	$ 56,851,043.00	73	89	$ 71,541,334.00
20	Detroit Tigers	$ 55,048,000.00	55	106	$ 49,356,167.00
21	Milwaukee Brewers	$ 50,287,833.00	56	106	$ 45,099,333.00
22	Kansas City Royals	$ 47,257,000.00	62	100	$ 35,422,500.00
23	Cincinnati Reds	$ 45,050,390.00	78	84	$ 48,784,000.00
24	Pittsburgh Pirates	$ 42,323,599.00	72	89	$ 57,760,833.00
25	Florida Marlins	$ 41,979,917.00	79	83	$ 35,562,500.00
26	San Diego Padres	$ 41,425,000.00	66	96	$ 38,882,833.00
27	Minnesota Twins	$ 40,225,000.00	94	67	$ 24,130,000.00
28	Oakland Athletics	$ 40,004,167.00	103	59	$ 33,810,750.00
29	Montreal Expos	$ 38,670,500.00	83	79	$ 34,849,500.00
30	Tampa Bay Devil Rays	$ 34,380,000.00	55	106	$ 56,980,000.00

Table 2. The 2002 payrolls by the MLB team

As an underdog team with a lack of star powers, the 2002 Oakland A's finished with a record of 103-59, the 2nd best in the 2002 MLB season. The team also held historically one of the longest winning streaks of 20 games during that season. The sabermetrics utilized in identifying players was the statistical analysis of data in baseball which aims to quantify baseball players' performances based on objective statistical measurements, especially in opposition to many of the established statistics that give less accurate approximations of individual efficacy (Neyer, 2017). For small teams like the Oakland A's, it is even more important to utilize analytics to identify opportunities and inefficiencies as they try to compete with teams with more resources.

Three-point evolution in the NBA

The idea of sports analytics has been implemented for quite some time in modern NBA history. Since the early 1990s, Pat Riley, one of the most prominent figures in the NBA, has been a pivotal figure in bringing about the use of analytics in basketball. As the coach of the New York Knicks and later as the general manager of the Miami Heat, he would watch game footage differently and would use a metric called PER (Player Efficiency Rating) to evaluate players. This metric considered various factors such as closeouts, boxouts, charges taken, and overall player efficacy (Buford, 2022). He was also the very first person to advocate the value of incorporating more three-point shots in the offensive scheme, leaving a long-lasting imprint that drove the league to take more three-point attempts per game for the past decades.

As mentioned in the introduction, the three-point line was implemented in the 1979-1980 season, and it was 22 feet far from the corners and 23.75 feet far from the top of the key to the basket. The distances were adjusted through the years but were readjusted back to the original distances. Though worth more than the regular two-point shots, teams were not utilizing and launching the three-point shots like they are today. As shown in Table 3, before the 1990s, teams were just attempting less than 10 shots a game (Chung, 2019).

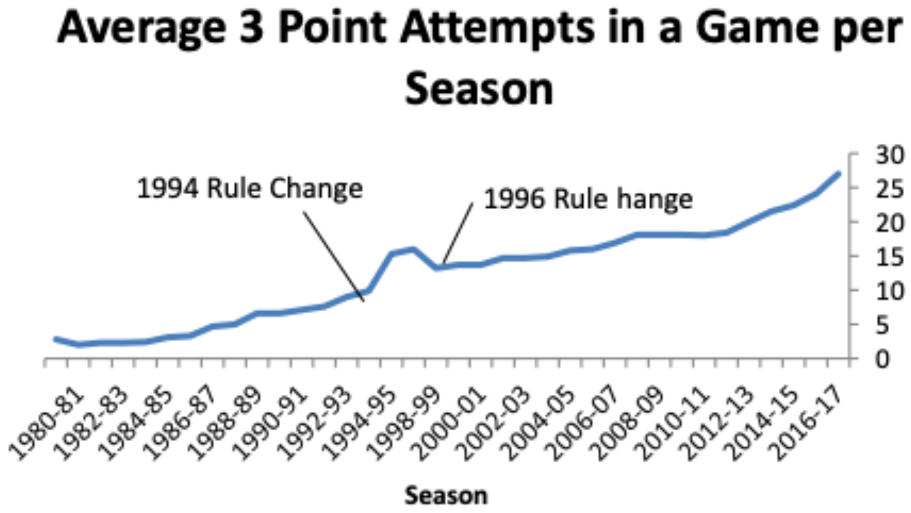


Table 3. Average three-point attempts in a game per season

The average attempts experienced a slight surge in 1994 when the league shortened the distance of the three-point line and later returned to normal when the three-point line returned to its original distance in 1996. Nevertheless, the adaptation has been continuous. In twenty years, the average three-point attempts in a game have reached 25 and continued to establish upward trends. The number of the three-point attempts of the whole league in the 2018 season also broke the NBA record with 78,732, 11.5 times higher compared with the first season when the three-point line was employed (Freitas, 2021). But what about three-point percentages? Again, in Chung's research (2019), it seems that, as shown in Table 4, the average three-point percentages have plateaued at 34% to 36% despite the continuous spike in three-point attempts. To make matters even more interesting, Chung also wished to discover whether the number of players that are considered "great shooters" has also increased. As presented in Table 5, the number of players that are capable of shooting over 36% continues to increase.

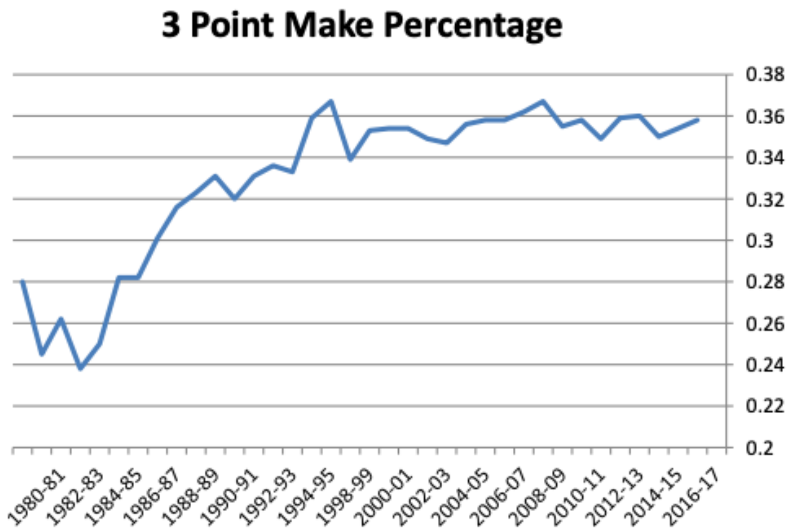


Table 4. three-point made percentages

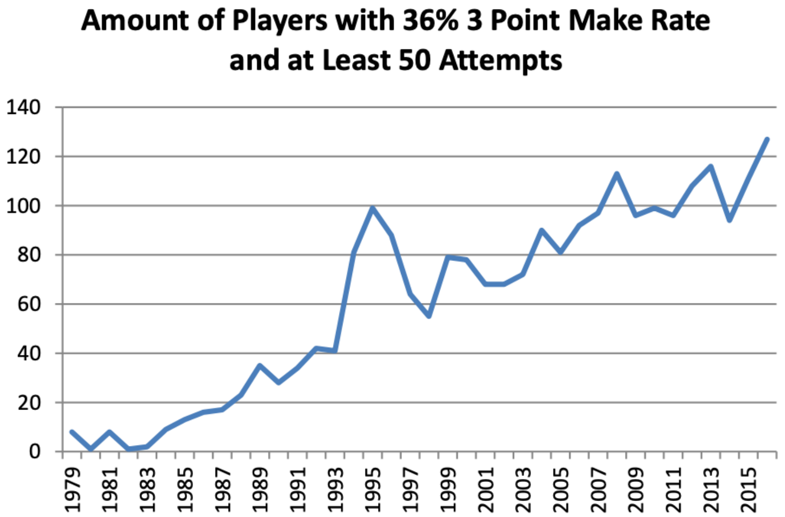


Table 5. Number of players with 36% 3 points make rate and at least 50 attempts

Just how obsessive are NBA teams with three-point shots? While it is intuitive to say that the field goal percentages decrease as the distances increase, it is also important to justify the shot selection by how much each type of shot is worth. According to Goldsberry's finding, by examining the 2017-2018 season, players shot an average of 39.6 % collectively from eight food, while shooting almost 36 % from the three-point line (2019), demonstrated in Table 6. The difference was almost marginal, suggesting that shooting long-distance two-point shots was almost idiotic.

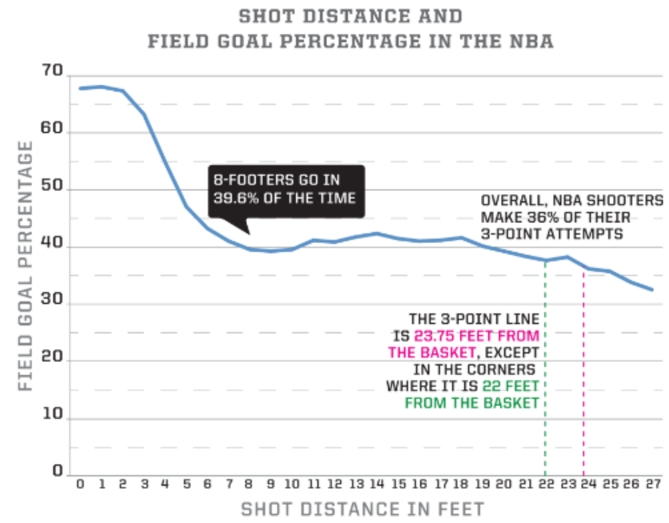


Table 6. Shot Distance and Field Goal Percentage in the 2017-2018 NBA Season

In addition, as suggested in Table 7, on average in the three-season span across 2013 – 2016, the NBA three was worth 1.07 points per shot, while a five-foot shot was worth 0.94 points, a six-foot shot was worth 0.87, and a seven-foot shot was worth 0.82. Again, suggesting shooting long-range two-point shots were just not that efficient.

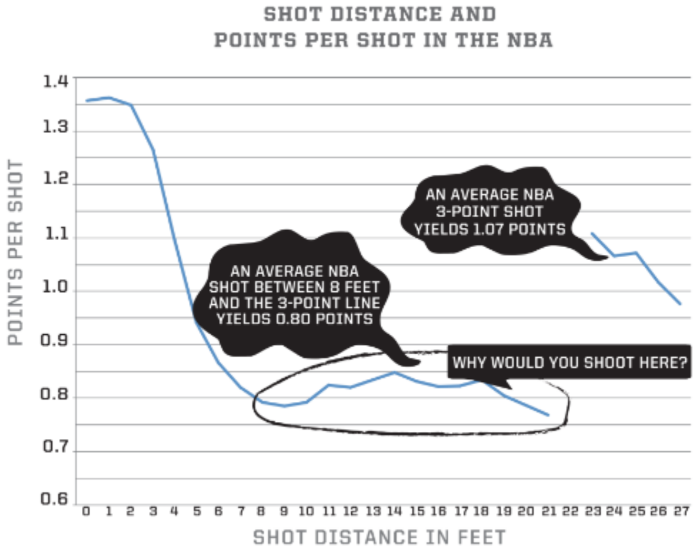


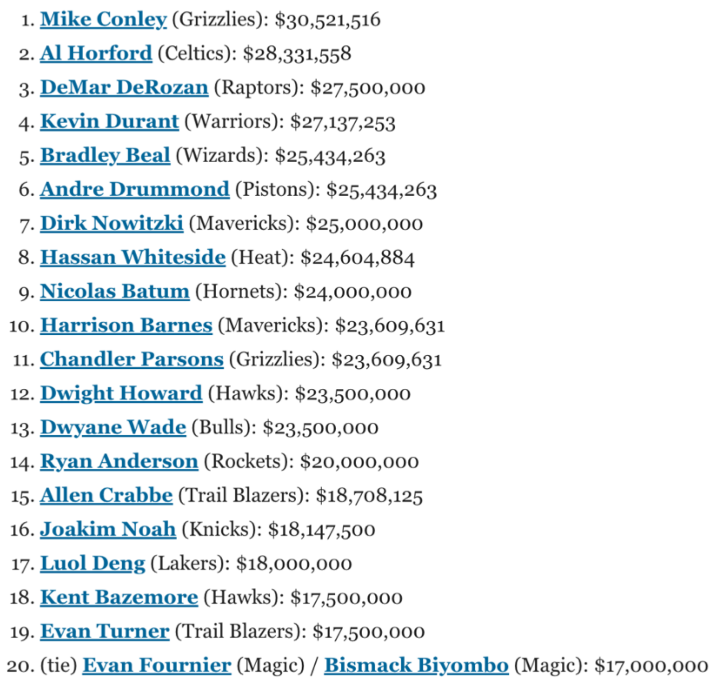
Table 7. Shot Distance and Points Per Shot in the NBA

<u>The new era of player categorizations</u>

To elaborate more on the wide adaptation of the three-point shots, let us focus on the 2000s, the golden age of three-point shots. The 2004 Phoenix Suns, headed by coach Mike D’Antoni, led the league in possessions per game, thanks to the offensive philosophy that suggested more possessions and faster-paced offense resulted in more fastbreak points and transition 3-pointers. The up-tempo style of basketball revolutionized the league, resulting in the majority of the teams in the league implementing faster-paced offensive schemes with quick three-point shots being taken. The evolutionary approach allowed players like Stephen Curry, smaller players who have mobility and sharp shooting ability to thrive in the new style of play. Stephen Curry proceeded to break the NBA record for the most 3-pointers made in a game and most 3-pointers made in a season during the 2014-2015 season. Curry's team, the Golden States Warriors, also established a long-lasting basketball dynasty specialized in the "run and gun" style of play in the 2010s. Spacing and implementing as many three-point shots in the offense ultimately gave the Golden States Warriors a successful championship run, resulting in the team reaching the NBA Finals five times and winning three titles from 2014 to 2019.

Such new paradigm shifts intrigued Daryl Morey, the president of basketball operations of the Philadelphia 76ers, and the general manager for the Houston Rockets between 2007 and 2020, who also happened to hold an MBA from Massachusetts Institute of Technology (MIT) and co-chaired the school's annual Sports Analytics Conference. NBA is a copycat league. After seeing the success of the Golden States Warriors, during his tenure with the Houston Rockets, Morey went all in on the principle of volume shooting from the three-point lines by surrounding James Harden, one of the most prolific scorers in the modern NBA, with a handful of shooters, hoping that Harden's ability to break down the defense would lead to as many open three-point shots as possible. The strategy ultimately led the team to construct a roster of players with shooting abilities, including bigs that were able to stretch the floor.

Ryan Anderson, who was drafted 21st in the 2008 NBA draft, was not a top talent in his draft class, nor was he a prolific scorer or defender. He started his NBA journey as a power forward for the Orlando Magic and The New Orleans Pelicans. However, he was awarded an $80 million four-year contract to join Morey's Houston Rockets in 2016. To put things into perspective, in Table 8, if we look at the top 20 free agent contracts of 2016 in terms of average annual salary, there were only 13 players that made more money than Anderson, and you could find several NBA all-stars, league MVPs, Olympic gold medalists, and future hall of fame players amongst the group (Adams, 2016). On top of it, Anderson, a power forward who only averaged 17 points and 6 rebounds in the previous season with the New Orleans Pelicans, didn't particularly stand out from a traditional sense in his position, yet he was one of the top earners in the NBA. Why is that?



1. **Mike Conley** (Grizzlies): $30,521,516
2. **Al Horford** (Celtics): $28,331,558
3. **DeMar DeRozan** (Raptors): $27,500,000
4. **Kevin Durant** (Warriors): $27,137,253
5. **Bradley Beal** (Wizards): $25,434,263
6. **Andre Drummond** (Pistons): $25,434,263
7. **Dirk Nowitzki** (Mavericks): $25,000,000
8. **Hassan Whiteside** (Heat): $24,604,884
9. **Nicolas Batum** (Hornets): $24,000,000
10. **Harrison Barnes** (Mavericks): $23,609,631
11. **Chandler Parsons** (Grizzlies): $23,609,631
12. **Dwight Howard** (Hawks): $23,500,000
13. **Dwyane Wade** (Bulls): $23,500,000
14. **Ryan Anderson** (Rockets): $20,000,000
15. **Allen Crabbe** (Trail Blazers): $18,708,125
16. **Joakim Noah** (Knicks): $18,147,500
17. **Luol Deng** (Lakers): $18,000,000
18. **Kent Bazemore** (Hawks): $17,500,000
19. **Evan Turner** (Trail Blazers): $17,500,000
20. (tie) **Evan Fournier** (Magic) / **Bismack Biyombo** (Magic): $17,000,000

Table 8. NBA’s top 20 free agent contracts of 2016 in terms of annual average salary

Turned out, Anderson was remembered as one of the defining “stretch-bigs” of the three-point generation (Goldsberg, 2019). Apart from setting screens and shooting the ball, the 6’10” forward was tasked to space the floor and to draw defenders or rim protectors to come out to guard the three-point line, drastically diminishing the defensive impacts of the opposing teams. With the newly implemented principle, the Houston Rockets ultimately made the deepest playoff run in the following season, ironically losing to the Golden States Warriors, the team that they modeled after, in the western conference finals.

Prolific guards that can shoot threes and stretch bigs are just two of the many new types of players in the modern NBA. Basketball fans can still see the five traditional position labels when watching a game, but the five positions are no longer suitable to represent the new breeds of players. Different points of view on how to redefine the positions were being introduced. In Bianchi, Facchinetti, and Zuccolotto’s journal, they suggested that there are five new positions in the modern NBA (2017), which are all-around all-star, scoring backcourt, scoring rebounder, paint protector, and role player. Another research conducted by Alaggapan suggested that the players in the league could be classified into thirteen new positions (2012), and they are offensive ball-handler, defensive ball-handler, combo ball-handler, shooting ball-handler, role-playing ball handler, 3-Point Rebounder, scoring rebounder, paint protector, scoring paint protector, role player, NBA 1st-Team, NBA 2nd-Team, and one-of-a-kind player.

Is it possible to predict the league's most valuable player (MVP) with game statistics?

One of the greatest debates that have baffled the basketball community would be the qualifications for the NBA's most valuable player honor. It is difficult to determine whether to choose individual players with top-notch statistical performances but with a losing record or great players with better winning records with less dominating in-game statistics. However, it is only fair to learn about the origin of the MVP award, as well as how the MVPs were chosen.

The MVP award is one of the most prestigious awards that an NBA player can receive during his playing career. It is decided by a committee of sports journalists and broadcasters in North America, and each member of the committee gets to cast votes for five players, while fans can also cast vote. The winner of the award is usually announced at the end of the regular season, sometimes around the first round of the NBA playoff. Though it is deemed paramount to players, history shows that the award qualifications are often bendable, sometimes valuing individual excellence, popularity, or how impactful an individual player is to change the dynamic of the team and led the team to win records. If the research was to settle the debate of what qualifies a player for the award with statistical modeling, just how would that turn out analytically?

To start, though the scope is slightly different, a hybrid model that combines Artificial Neural Network method and other supervised learning algorithms proposed by Albert, Lopez, Allbright, and Blas in 2021 successfully backtested and selected the 1998 NBA all-stars players with 88.7% accuracy, indicating statistics-based award selections in the NBA is plausible. What about the MVP awards?

Many studies have been conducted throughout the years on this matter. In Mason Chen's research from 2017, he extracted performance data from the 2003 to 2016 season and used four statistical models, the Uniform Model, the Weighted Model, the Power Model, and the Discriminant Clustering Model, to predict the 2016-2017 season NBA MVP. The Uniform Mode, which served as the foundation of other proposed models, was improved, and optimized by applying factors, aiming to achieve better accuracies. Among all proposed models, the Power Model successfully predicted 69% of the time, suggesting the selection of NBA MVP of each season can be more transparent and objective, whether it be voted by the media partners or by analyzing the performances.

Finally, a Back Propagation Neural Network algorithm was developed by Hu, Zhang, and Qiu in 2019 to test whether the model could pick out NBA MVPs from 2010 to 2018. They claimed that the Neural network model has the capability to learn and create models for complex, non-linear relationships that serve as the foundation of MVP prediction model. When constructing one, it's important to make necessary adjustments based on the specific scenario to obtain results that are more accurate, satisfactory, and realistic.

III. Methodology

Datasets

To thoroughly capture the entirety of the research findings and the essence of sports analytics, multiple sources of data are gathered and analyzed to create a comprehensive and complete understanding.

Curated by Sumitro Datta regularly up till 2023 on Kaggle, a significant portion of the data gathered includes information from three different leagues: the National Basketball Association (from 1950 to the present), the Basketball Association of America (from 1947 to 1949), and the American Basketball Association (from 1968 to 1976). To make it easy to gather career statistics for each player, each player was assigned a unique ID. Moreover, the data collection includes 7 team-related datasets, which provide totals, per-game statistics, and per-100 possessions statistics (starting from 1974) as well as team summaries. On the player side, 10 datasets offer a range of information such as player totals, per-game stats, per 36-minute stats, per-100 possessions stats (starting from 1974), advanced stats, play-by-play stats (starting from 1997), shooting stats (starting from 1997), end of season team selections (All-Defense, All-Rookie, All-League), end of season team voting (All-League), all-star selections, and awards voting results (Rookie of the Year, Sixth Man of the Year, Most Valuable Player, Defensive Player of the Year, Most Improved Player).

On the other hand, the data collection gathered by Nathan Lauga on Kaggle is updated frequently every quarter. It includes 5 datasets, such as the game data which contains all games from the 2004 season to the most recent update, with information such as the date, teams, box scores, opponents' box scores, and the record of wins; games details data which provides in-game detailed statistics of every player for a given game; players data which has information on the players such as name, experience, age, team, and more; ranking data which shows the ranking of NBA teams on a given day, splitting into the west and east conference; and the teams' data which has information on all the teams in the NBA.

Lastly, to identify the impact of shot locations, the dataset curated by Sports Viz Sunday on Data.world website includes a wealth of information directly from the NBA Stat API, including every shot location from the 1997-98 season up to the present day, totaling almost 5 million shots. The shot information that comes with every shot includes the player who took the shot, which team did the player belong to, the time of the shot, the type of shot taken, the location of the shot in distances (ex. Feet), and in generality (ex. Left Corner 3), and whether the shot has been made or not.

Methods to identify three-point shot trends and patterns

Firstly, by using exploratory analysis to identify patterns and trends in datasets, several inferences shall be reached and identified to validate the existence of three-point evolution. To do so, datasets that include teams' and players' total statistics of each season such as games played, minutes played, numbers of field goals made and attempted, free throws, two-point, and three-point field goals made and attempted, and other peripheral stats (ex. rebounds, assists, steals, blocks, turnovers), and the shot location dataset were explored and analyzed. In the analysis, only the information after 1979 when the three-point line was implemented was examined.

After the dataset is combined and cleaned, the first essential metrics to visualize are the average number of three-point attempts and the average three-point field goal percentages throughout different seasons to determine if there's a significant trend in implementing the three-point shots in the game. Once this is computed, we should be able to see whether there's a progressive upward trend in both attempts and shots made through different eras.

Not only are there signs of wide adaptation, but to examine whether shooting performances also improve on a league-wide level, the average three-point shot field goal percentages as well as the standard deviation are calculated. To support the assertion that there is a significant uptrend in the adaptation of three-point shots at the league level, the number, and the ratio of elite three-point shooters within the league, the correlation between shot distance and field goal percentage, and the correlation between shot distance and points per shot are to be presented. These metrics should provide sufficient statistical evidence for the benefits of taking more three-point shots over two-point shots. This amount of information should effectively demonstrate and validate the reason behind the ongoing evolution towards a greater emphasis on three-point shots in the league.

Methods to classify NBA players into natural groupings with box scores

The next area of inquiry is determining how to re-conceptualize player positions in modern basketball. Since the NBA collects a significant amount of data, large datasets often contain a high number of dimensions and features for every observation, not to mention features that may be highly correlated. Instead of assigning players to traditional positions based on physical attributes or skill sets, the objective is to use in-game statistics to group players into categories. To achieve this, two unsupervised learning techniques, the K-Means clustering algorithm, and the Hierarchical clustering algorithm are employed to uncover patterns within the data, as opposed to supervised learning methods that rely on pre-existing position labels or target values.

Some of the reasons that the K-Means algorithm is preferred in this specific research topic due to its ease of use and implementation, and its ability to achieve high computational efficiency. K-means assumes that all features have equal variances, so if the dataset does not meet such pre-requisite, the algorithm may not produce meaningful results. Thus, performing a variance test before running k-means can help identify any features with vastly different variances and allow standardization to happen before clustering. By using k-means, the players will be grouped into k clusters, where k is a user-specified number of clusters. Each player will be assigned to the cluster with the nearest mean. Once the clusters have been created and labeled, the players within each cluster can be analyzed to see if there are any patterns or similarities among the group. This can help understand which players have similar statistical profiles and how they differ from players in other clusters.

On the other hand, the Hierarchical clustering algorithm is utilized because it not only does not require the number of player clusters to be specified in advance like the K-Means method, but it is commonly used to identify natural groupings or patterns, instead of classifying data into distinct player categories. The Hierarchical clustering algorithm produces a dendrogram, which shows the hierarchical relationship between the player clusters. Once the optimal number of player clusters is determined by using the selected linkage method parameter that provides the highest Silhouette score, the player clusters are formed, and the average statistics of each player cluster are also presented for more justification.

Methods to predict NBA MVPs

The last challenge is to identify a data-driven method that helps to excel in the MVP selection processes more statistically. In contrast to the previous task in which no specific values were being targeted, in this challenge, various supervised learning models are trained using labeled data, specifically, the MVP winners from the past several decades, to make predictions about MVP. To achieve this, all MVP candidates' average statistics data, award winners, and team records from 1980 to 2015 are used to train the data, whereas the data from 2016 to 2022 are used to test the feasibility of the trained models.

Three supervised learning models are built and compared so that ultimately a model with the highest accuracy and the lowest error may be identified for future reference. Before introducing the algorithms, the dataset is tested for multicollinearity and analyzed with PCA at certain stages of the analysis to reduce complexity and avoid overfitting for model building. The models introduced later are also fitted twice, once before, and once after implementing correlation and PCA analyses to explore and compare the impacts of multicollinearity and dimensionality reduction.

The three methods implemented in this challenge are Support Vector Machine, Decision Tree, and Random Forest. Support Vector Machine is selected due to its effectiveness in high dimensional spaces like the dataset that is used in the analysis, on top of it being able to handle non-linear decision boundaries in case the dataset is linearly non-separable. However, it is computationally quite expensive and is sensitive to the choice of kernel and parameters. In this model, three parameters in the regularization parameter (C), kernel type, and gamma are tested. The regularization parameter in the SVM algorithm is used to balance the trade-off between classification error and the distance between the decision boundary and the closest data points. A higher value of the regularization parameter leads to a higher variance and lower bias, while a lower value leads to lower variances and higher biases. The values of the regularization parameter are tested with different numbers, such as 10, 100, 1000, 10000, and 100000. The algorithm uses different types of mathematical functions called kernels, such as Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid, to transform the input data into a suitable form for model building. The goal is to identify a kernel method that produces the highest accuracy and the lowest mean squared error. Additionally, there is a parameter called gamma, which is the coefficient for the kernels. It controls the influence of individual training samples. The gamma values are tested with different numbers, such as 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, and 100000.

Next, the Decision Tree method is easy to comprehend and interpret and can handle categorical and numerical data, missing values, and outliers quite well. But small variations in the dataset might cause the model to overfit. The parameter the maximum number of leaf nodes (max_leaf_nodes) is tested, which represents the maximum number of decisions or predictions made by the tree. The fewer the number of leaf nodes, the more simplified the model, though it may also have higher biases. In other words, it is prone to overfitting when the tree becomes too complex.

The algorithm that builds on the foundation of the Decision Tree algorithm is the Random Forest algorithm. Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their predictions. The algorithm has two main parameters to be tested in this research challenge, the number of trees in the forest (n_estimators) and the number of features to be considered when looking for the best split (max_features). The values of n_estimators are tested with different numbers, such as 10, 100, 1000, and 10000, while the max_features parameter is tested with numbers ranging from 1 to 10. The final decision is made by taking a majority vote among the trees in the forest. Algorithms such as Gradient Boosting and Neural Networks both are known for making accurate predictions, yet these two algorithms require even higher computational powers, can be sensitive to the choice of hyperparameters, and require a lot of tuning to get good results, thus they unfortunately are not included in this research analysis.

IV. Analysis

Three-point shooting phenomenon analysis

As suggested in the Methodology section, the Teams’ Total and Players’ Total datasets containing in-game statistics since 1979 are examined to check for evidence for three-point shooting trends.

Firstly, the average number of three-point shots attempted and made are computed by dividing the total number of three-point shots taken and made by the league by the total number of games played for each season. The two metrics are then plotted against the season feature to check for any trends that are worth mentioning. Next, the mean of the three-point shots field goal percentage and standard deviation are calculated to determine the shooting averages of the league, and how much variation or "dispersion" there is in the three-point shooting performances over the years. To determine if the high number of three-point shots taken is a trend across the entire league and not just limited to a few players, we need to examine the proportion of elite shooters in the league. In this research analysis, the elite shooters are those that has made at least 50 shots and shooting at least 36% from the three-point line.

Lastly, the relationship between shots distances and the field goal percentage associated with the distance, and the relationship between shots distances and the average points per shot associated with the distance are plotted. These two metrics are designed to identify the effectiveness of a three-point shot, and the risk-reward ratio of taking a three-point shot versus taking a two-point shot.

NBA Player positions reimagining

When evaluating a player's career, it is common to look at their career averages. In this research analysis, to classify players statistically, the dataset used includes information such as the player's birthday, position, team, and number of games played. It also includes box scores for each season, including statistics like field goal and free throw attempted, made, field goal and free throw percentages, rebounds, assists, steals, blocks, turnovers, personal fouls, and points scored. After removing all non-numerical values and checking for missing values, to calculate a player's career averages, the total box scores from each season before 1989 are added together and divided by the number of games played. The player statistics included in the final dataset to be used to build the models for categorization are as follows:

- player_id (Player ID)
- player (Player name)
- mpg (Minutes per game)
- fg_pg (Field goal made)
- fga_pg (Field goal attempts)
- fg_percent (Field Goal percentage)
- x3p_pg (Three-point made)
- x3pa_pg (Three-point attempts)

- x3p_percent (Three-point percentage)
- ft_pg (Free throw made)
- fta_pg (Free throw attempts)
- ft_percent Free throw percentage)
- ppg (Points per game)
- trpg (Total rebounds per game)
- orpg (Offensive rebounds per game)
- drpg (Defensive rebounds per game)
- apg (Assists per game)
- spg (Steals per game)
- bpg (Blocks per game)
- tpg (Turnovers per game)
- pfpg (Personal fouls per game)
- season (Seasons played)

To briefly mention the utilized dataset, there are no specific outliers that stand out, which is particularly advantageous since the K-Means algorithm does not work particularly well with outliers presented compared to other unsupervised learning algorithms. The dataset has a total of 2,508 observations and 22 features in total, in which the domain for minutes per game is from 0 to 43.5, whereas for three-point field goal percentage is from around 0 to 1. With such significant differences, to avoid overfitting, improve the quality of clusters, and increase the accuracy of clustering algorithms, after removing non-numerical attributes such as player ID, player names, position names, variance test, and standardization are performed to adjust the variability of the final dataset by using a linear transformation to convert data into a specific range between 1 and -1. The features in the dataset are to be ensured that the variances are in a similar range and are less than 0.05 after performing standardization, which is beneficial when performing K-Means Clustering analysis.

To increase interpretability and reduce dimensionality, Principal Components Analysis (PCA) is utilized to transform the data in ways that help identify the most important underlying structure in the data and represent it in a new set of derived variables, called principal components. The principal component is the linear combinations of the original variables, and within each component is a handful of top player features that contribute the most to the principal component, or in this case, the features that have the most impact on the player clustering algorithm. To understand a large dataset with the least amount of information, the PCA analysis groups the data into smaller groups called principal components. These groups are made in a way that represents the most about the larger dataset. The PCA analysis then shows how much each component explains about the overall player statistics, which is called explained variance. A good rule to follow is to use all groups that explain more than 80% of what is going on in the larger dataset accumulatively.

Once the number of principal components is determined, two common practices to determine the number of player clusters are used, and they are the elbow method and the silhouette method. The elbow method identifies the optimal number of clusters where the within-cluster sum of square, or the measure of the similarity of the data points within the cluster, starts to decrease at a slower rate. In simpler terms, the method looks for an optimal number of player categorizations where adding more groups does not increase each player cluster’s measure of similarity as much. On the other hand, the silhouette method identifies the best number of player categorizations where each piece of data is most like the data in its own player position and the least like the data in others. A high silhouette score means that each piece of data is well-matched to the group it belongs to and not well-matched to the other groups. A score above 0.7 is usually considered high, indicating a good grouping of data.

After the number of player positions is determined, the next steps are to fit the model, with the PCA scores, to apply PCA scores to each player, and to consequently assign player clusters to each observation. Once the clusters are labeled, the average statistics of each cluster are computed, and the visualization of players' clusters concerning the PCA components is plotted. By doing so, such a plot can help determine whether the new player positions assigned by the K-Means clustering algorithm are successfully categorized in which all clusters are visually separated from each other.

Next, an important step to improve the performance of the clustering algorithm is by using a dimensionality reduction technique like PCA before data segmentation to decrease the number of features and to reduce noise. After fitting the PCA variable with standardized data, it generates as many principal components as there are features in the data, in this case, 22, arranged in the order of importance, or explained variance. The captured amount of explained variance and the amount of cumulative explained variance depending on the number of components are displayed. Since the goal of PCA is to retain as much of the variation in the data as possible while still reducing the dimensionality, this research aims to retain at least 80% of the explained variance. Consequently, 4 principal components are retained.

Next, the Elbow method and the Silhouette Score methods are put in place to determine the number of clusters when performing the K-Means algorithm. The within-cluster sum of square (WCSS) and the Silhouette scores are each placed against a range of cluster numbers from 1 to 15. Once the optimal number is determined, the K-Means clustering model is fitted with the PCA scores from above, and the original player dataset with designated PCA scores and the cluster numbers are labeled after each player. The players are then grouped into clusters for comparison.

Similarly, the Hierarchical Clustering technique is then performed to see whether other unsupervised clustering techniques produce a similar result. Though the number of optimal player categorizations required is also determined by the Silhouette method, the other two parameters that are being tested in this method are the linkage method and the criterion, which is used in forming flat clusters, whereas linkage helps to identify the maximum distances between each pair of clusters. The ward linkage method is chosen because it minimizes the total within-cluster variance and can handle categorical data well. On the other hand, the Maxclust and the Distance criteria are also tested to see how the clusters are formed in the Hierarchical Clustering algorithm. The players are then again grouped into clusters based on the number of clusters suggested for comparison.

NBA MVP prediction analysis

To train the MVP prediction model with the information required, below are the data included in the final MVP candidates' dataset:

- season (Season)
- player (MVP candidates' name)
- age (Age)
- tm (Team Abbrev. Eg. HOU)
- pts_per_game (Points per game
- mp_per_game (Minutes per game)
- fg_per_game (Field goal made)
- fga_per_game (Field goal attempts)
- fg_percent (Field Goal percentage)
- x3p_per_game (3-pts made)
- x3pa_per_game (3-pts attempts)
- x3p_percent (3-pts percentage)
- ft_per_game (Free throw made)
- fta_per_game (Free throw attempts)
- ft_percent (Free throw percentage)
- trb_per_game (Total rebounds per game)
- orb_per_game (Offensive rebounds per game)
- drb_per_game (Defensive rebounds per game)
- ast_per_game (Assists per game)
- stl_per_game (Steals per game)
- blk_per_game (Blocks per game)
- tov_per_game (Turnovers per game)
- pf_per_game (Personal fouls per game)
- per (Player efficiency ratings)
- usg_percent (Usage percentages)
- w (Teams wins)
- l (Team losses)
- mov (Margin of victories)
- winner (MVP award. Yes: 1, No: 0)

As suggested in the Methodology section where the multiple tabular datasets are merged and scaled after cleaning and transforming, and the remaining final dataset includes all MVP candidates' player statistics from 1980 to 2022, winning records of the teams the players played for, and the dummy variable where the player who won the MVP is labeled 1, and 0 otherwise. The non-numerical values are first removed to check the multicollinearity. The correlation coefficients between all features are then shown. A high number of features in the dataset that are highly correlated to each other suggests multicollinearity, which may influence the final prediction.

Instead of the usual practice of splitting the dataset randomly into a 70% training set and 30% testing set, the dataset is split manually into two groups where the training set is the MVP candidates' data before 2015, and the testing set is the MVP candidates' data after 2016. In Table 9, the information for the league MVP since 2016 is shown. The ratio of such data split is about 84:16. Lastly, the feature "mvp" is further separated to serve as the target values. Once the dataset is split and the non-numerical features removed, both training and testing datasets are standardized and converted into a specific range between 1 and -1.

Season	Award	Player	Age	Team
2022	MVP	Nikola Jokić	26	Denver Nuggets
2021	MVP	Nikola Jokić	25	Denver Nuggets
2020	MVP	Giannis Antetokounmpo	25	Milwaukie Bucks
2019	MVP	Giannis Antetokounmpo	24	Milwaukie Bucks
2018	MVP	James Harden	28	Houston Rockets
2017	MVP	Russell Westbrook	28	Oklahoma City Thunder
2016	MVP	Stephen Curry	27	Golden States Warriors

Table 9. MVP winners since 2016

From here, the standardized dataset that is yet to be put through PCA transformation is put through the Support Vector Machine, Decision Tree, and Random Forest algorithms to build the first group of models. As mentioned in the Methodology section, each parameter of all three algorithms is tested with ranges of values through nested loops to find the optimal combinations of parameters that provide the best results. Once the models are developed, metrics such as accuracies, the performance metrics associated with the confusion matrix (True Positive Rate, Precision, Recall, False Negative Rate, and F1-Score), and the mean squared error (MSE) are plotted to compare the three models' performances and feasibilities.

Next, the standardization data is transformed with PCA analysis to identify the optimal numbers of principal components that explained at least 80% of the variance in the dataset. The selected components are then proceeded to check for multicollinearity again to ensure highly correlated components are eliminated. Support Vector Machine, Decision Tree, and Random Forest algorithms are then trained with the PCA transformed dataset, and the performance metrics of the new group of models are presented.

In short, all three models are trained before and after the multicollinearity and dimensionality reductions, giving us six models in total for comparison.

V. Discussion

The evidence of a league-wide three-point shots evolution

By plotting the average number of three-point shots attempted and made as displayed in Table 10, several phenomena may be observed.

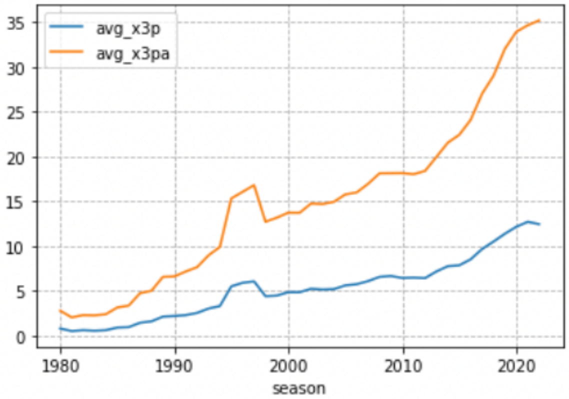


Table 10. Average three-point shots made and attempted since 1980

From the table, it is quite evident that the league did not burst into launching as many three-point shots in the beginning until mid-1990s. As suggested in the introduction, the three-point line was shortened in 1994 by the league, hoping such a change would encourage teams to take more three-point shots for entertainment purposes, which was demonstrated by the deep surge in both shot attempts and made in the chart. Still, by that time, the league was only taking an average of ten three-point shot attempts and making only an average of four shots. The numbers continued to rise and saw exponential growth in both shot attempts and made between 2012 and 2020. The inclination of the incremental increase in both shot attempts and shot made coincides with Chung's findings in 2019, suggesting the attempts had increased to 25 per-game, and continued to rise, as demonstrated in Table 3. By the end of 2020, the league was taking an average of thirty-four three-point attempts, while making an average of twelve of them. It is quite remarkable to see how the league has significantly improved in attempting and making three times as many three-point shots in just 30 years, compared to 1994.

In the same findings, Chung suggested that the average three-point shooting percentage plateaued at around 2000. In Table 11 where the average three-point shots field goal percentages are plotted, it is noticeable that before 2000, the league had displayed turbulent three-point shooting performances, shooting at as low as 24%, and as high as 36%. Since 2000, the performances have stabilized and plateaued at 34% to 36%, which coincides with Chung's research in 2019. On top of it, to demonstrate how the teams in the league were not yet comfortable with the three-point shots before the 2000s, Table 11 also shows the standard deviation of the three-point field goal percentages for each season. While the shooting percentage continued to rise and stabilize, the standard deviation continued to diminish from 0.05 in the 1980s to nearly 0.02 in the 2000s. To further suggest how progressive the league is, as suggested by Chung, the definition of an elite shooter is someone who attempts at least 50 three-point shots in a single season while shooting at least 36%. In Table 12, the chart indicates that this type of player has made up more than 25% of the roster spots now in the 2020s, compared to just 3% back in 1980. The chart illustrates the increasing trend of teams in the league considering the use of more shooters on their roster, which is what caught Chung's attention as well.

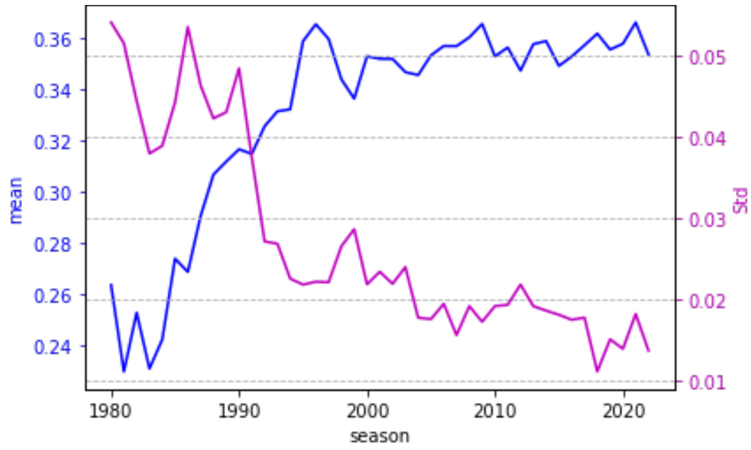


Table 11. Average three-point shots percentage and the standard deviation

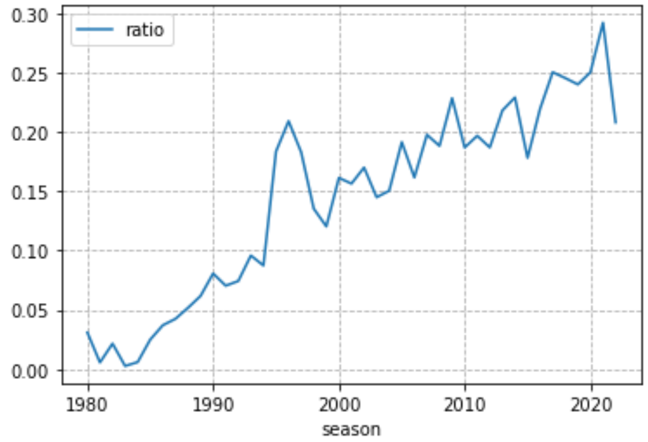


Table 12. The ratio of elite shooters in the league over the years

Now, to ingest the concept of sports analytics, according to Goldsberry's findings, there are interesting relationships between the shot distances and both field goal percentages and points per shot. Two of the reasons for teams launching more three-pointers, as suggested by Daryl Morey after implementing sports analytics in the NBA, are as follows. As seen in the previous charts, since the 2000s after the league-wide average three-point shot percentage plateaued, the number of attempts continued to rise regardless. In Table 13 where the shots distances and the average field goal percentages are plotted, the shooting percentages from the three-point line (where the red vertical lines are plotted) in 2000, 2010, and 2020 were comparable to those from 3-22 feet within the line, except for shots taken from 0-3 feet, which are typically layups or dunks. Moreover, in Table 14 where the shots distances and the average points of shot are plotted, the points per shot from the three-point line are simply worth more than from 3-22 feet. As suggested by Morey and Goldsberry in Tables 6 and 7, such a risk-reward ratio strongly implored NBA teams to launch more three-point shots. In short, it is undeniable that the above supporting facts strongly suggest the claim that there is a significant and continuous three-point adaptation in the NBA.

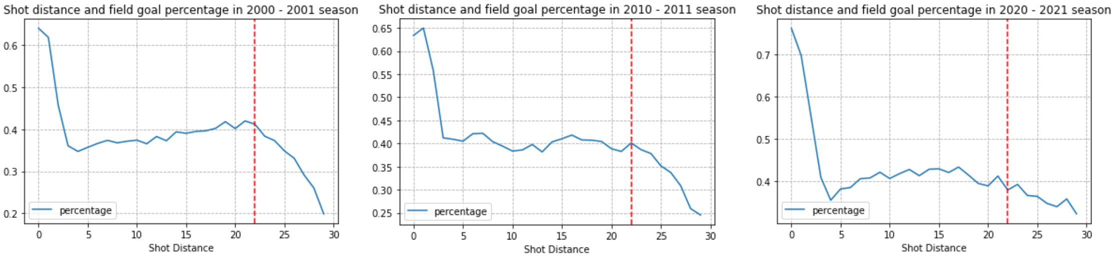


Table 13. Shot distance and field goal percentages in three different seasons

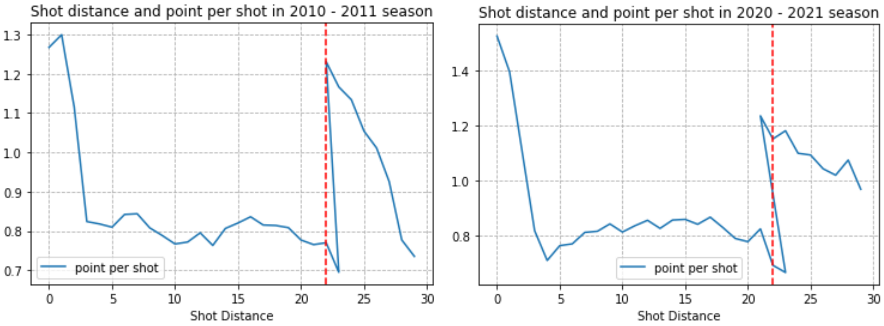


Table 14. Shot distance and points per shot

Modern players' positions classification

As suggested in the Analysis section and in Tables 15 and 16, 4 components are preserved as the cumulative explained variance is 82.69% when performing PCA analysis on the players' dataset. To elaborate on the preserved components, the top 5 features that contribute the most to principal component 1 are personal fouls per game, field goal and field goal attempts per game, points per game, and minutes per game. Features that contribute to component 2 are assists per game, free throw percentages, three-point percentages, three-point attempts and made per game. Features that contribute to component 3 are field goal attempts per game, turnover per game, free-throw made and attempts per game, and two-point field goal attempts per game. Features that contribute to component 4 are rebounds per game, blocks per game, defensive rebounds per game, and three-point shots made and attempts per game. Ultimately, it is only required to calculate the remaining 4 components' scores for the elements by fitting the PCA model again, so the K-Means clustering can be performed based on principal components scores instead of the original 24 features.

Component	Explained Variance	Cumulative Explained Variance
1	0.5661	0.5661
2	0.1517	0.7179
3	0.0586	0.7765
4	0.0504	**0.8269**
5	0.0321	0.8589
6	0.0282	0.8871
7	0.0252	0.9123
8	0.0214	0.9337
9	0.0172	0.9510
10	0.0127	0.9637
11	0.0107	0.9744
12	0.0101	0.9845
13	0.0052	0.9897
14	0.0040	0.9936
15	0.0032	0.9968
16	0.0022	0.9990
17	0.0005	0.9995
18	0.0003	0.9998
19	0.0002	1.0000
20	0.0000	1.0000
21	0.0000	1.0000
22	0.0000	1.0000

Table 15. Explained Variance and Cumulative Explained Variance of the PCA Components

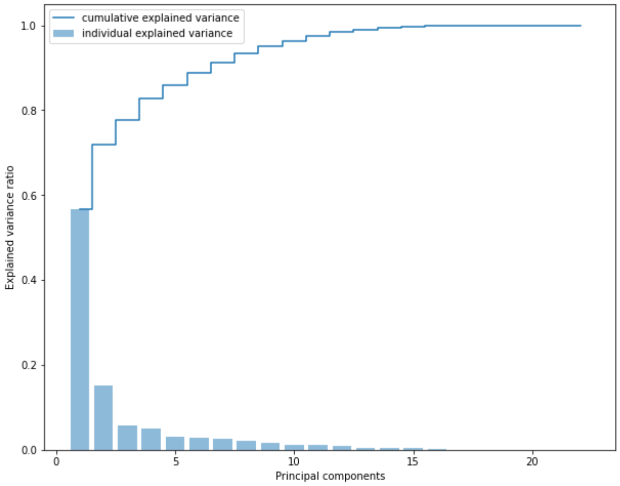


Table 16. Explained Variance by components

Moreover, in Table 17, the elbow method suggests that the "elbow" point in the plot is having 2 clusters where the within-cluster sum of square (WCSS) starts to decrease at a slower rate, indicating that adding more clusters will not significantly decrease the

WCSS. Moreover, the Silhouette method also suggests that having 2 clusters gives the highest silhouette coefficient values, indicating having 2 clusters should separate data points well.

Nevertheless, it would be quite pointless and bizarre to say that there are only two classifications of players since the 1980s, so the number of clusters that results in the second-highest average silhouette coefficient value is considered as the next best alternative after having 4 clusters. Next, by designating the number of clusters to 4 when fitting the K-Means clustering with the PCA scores from above, a dataset with designated PCA scores and the cluster numbers are labeled after each observation. After calculating the average of the in-game statistics of all players in each cluster, Table 18 shows the number of players that are assigned to each of the 4 clusters, and Table 19 illustrates the average statistics of the newly assigned 4 positions are displayed.

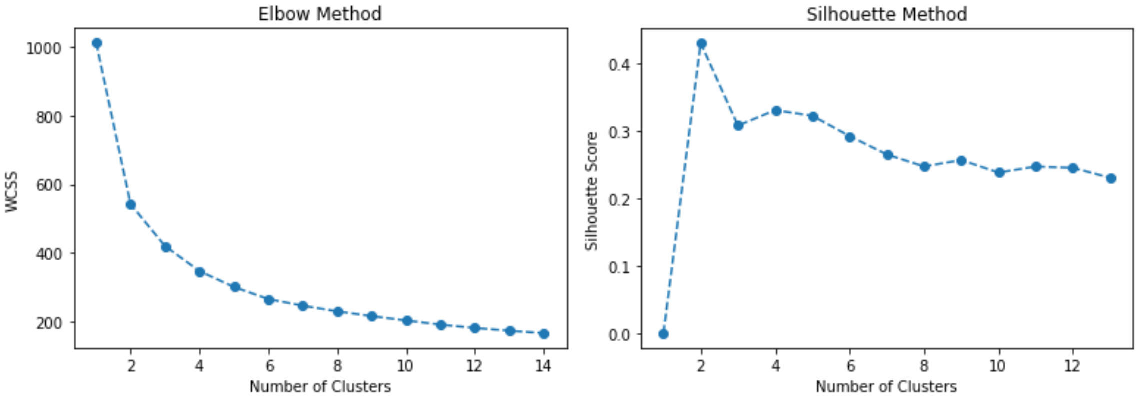


Table 17. Elbow Method & Silhouette Method

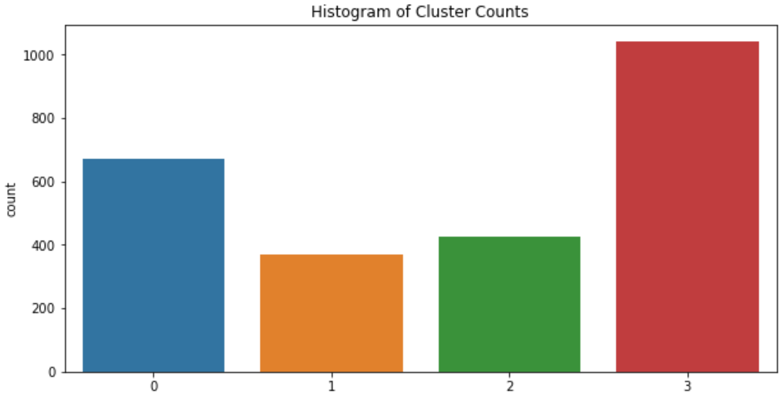


Table 18. Number of players in each classified cluster using the K-Means clustering method

cluster	mpg	fgpg	fgapg	x3ppg	x3papg	x2papg	ftpg	ftapg	ppg
1	20.72	2.87	6.73	0.87	2.46	4.27	1.16	1.5	7.78
2	30.97	5.77	12.61	1.06	3	9.61	3.03	3.87	15.64
3	20.26	2.94	5.97	0.08	0.29	5.68	1.41	2.06	7.38
4	10.01	1.2	3	0.23	0.79	2.21	0.56	0.81	3.19

cluster	fg_percent	x3p_percent	x2p_percent	ft_percent	trpg	orpg	drpg	apg	spg	bpg	tpg	pfpg	season
1	0.43	0.34	0.47	0.77	2.66	0.59	2.07	2.01	0.7	0.24	1.09	1.77	8.67
2	0.46	0.33	0.49	0.79	5.15	1.25	3.9	3.64	1.06	0.55	2.12	2.39	12.22
3	0.49	0.16	0.51	0.67	5.14	1.75	3.39	1.01	0.55	0.73	1.11	2.36	10.19
4	0.4	0.24	0.43	0.69	1.58	0.49	1.1	0.79	0.32	0.17	0.58	1.1	3.46

Table 19. Newly assigned player positions' statistics by K-Means clustering method

Like what Bianchi, Facchinetti, and Zuccolotto's journal suggested where there is an all around all-star, scoring backcourt, scoring rebounder, paint protector, and role players, in Table 19, there are four new player groups suggested by the K-Means clustering model, and they are all-around all-stars, three-point specialists, defensive big men, and role players. The scoring wing player of cluster 1 has the best three-point shooting capabilities and shoots an equal number of three-point shots as all-around all-stars. Famous sharpshooters like Dell Curry and Dennis Scott, to more modern players like Desmond Bane and JJ Redick all belong to this group. Ryan Anderson, the player that is mentioned in the introduction that sparks the debate of player re-categorization, also belongs to this cluster. Next, in cluster 2 are the all-around all-stars. These players play the most minutes, score the most points, have the most field goal attempts and made, and play the greatest number of seasons on average. Well-known players like Michael Jordan, Lebron James, Kevin Durant, Allen Iverson, Shaquille O'Neal, and many more franchise players in NBA history, all belong to this cluster. In cluster 3, defensive big men are usually the ones that specialize in rebounds and shot blocks. These players average the most blocks and rebounds, have the highest field goal percentage, and usually have the most offensive rebounds. Famous players like Charles Oakley, Dikembe Mutombo, Ben Wallace, Rudy Gobert, and the infamous Dennis Rodman all share the same traits. Lastly, cluster 4 consists of more functional players. These players usually don't stand out statistically, have short NBA careers, and are not the focal points of the offense. Interestingly enough, according to Table 18, this type of player makes up the most rosters spots throughout the years.

Lastly, in Table 20, the visualization of clusters concerning the first two PCA components is plotted. It is quite clear that each segment is separate clearly without much overlapping, suggesting the features in the first two components can create clusters that are distinct from each other.

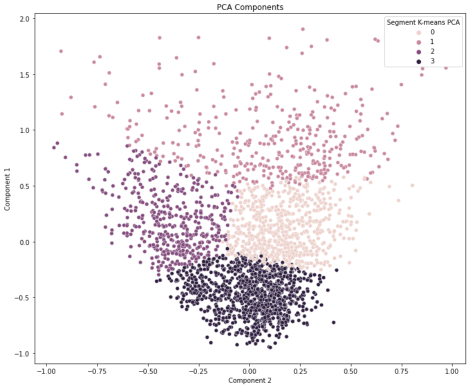


Table 20. Visualization of clusters concerning the first two PCA components

Next, As seen in Table 21, the maxclust criterion with the ward linkage method produced the second highest silhouette score at 0.25 with 6 clusters as we don’t consider 2 as the optimal number of clusters. After selecting Maxclust method as the criterion parameter, clusters are extracted and labeled for each observation based on the linkage matrix as shown in Table 22.

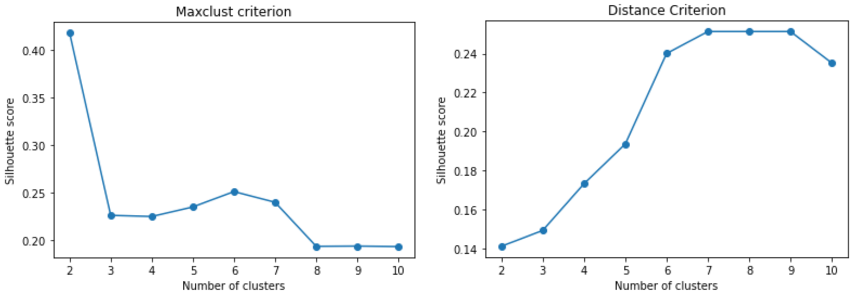


Table. 21 Silhouette scores of different numbers of clusters using different linkage methods

cluster	mpg	fgpg	fgapg	x3ppg	x3papg	x2papg	ftpg	ftapg	ppg
1	14.02	1.79	3.77	0.06	0.22	3.56	0.85	1.3	4.5
2	33.27	7.02	14.65	0.81	2.38	12.27	4.13	5.34	18.98
3	27.03	4.39	9.95	1.05	2.94	7.01	2.04	2.58	11.87
4	17.56	2.37	5.61	0.7	2.03	3.58	0.94	1.25	6.37
5	8.3	0.96	2.56	0.23	0.82	1.74	0.45	0.65	2.6
6	23.5	3.51	7.06	0.05	0.18	6.88	1.75	2.56	8.83

cluster	fg_percent	x3p_percent	x2p_percent	ft_percent	trpg	orpg	drpg	apg	spg	bpg	tpg	pfpg	season
1	0.48	0.14	0.49	0.65	3.31	1.17	2.14	0.67	0.4	0.45	0.76	1.87	6.93
2	0.48	0.29	0.51	0.78	7.28	1.87	5.41	3.78	1.07	0.92	2.52	2.65	11.78

3	0.44	0.34	0.48	0.79	3.63	0.83	2.8	3.11	0.95	0.33	1.64	2.12	12.27
4	0.42	0.33	0.47	0.76	2.37	0.57	1.8	1.6	0.59	0.23	0.92	1.59	5.74
5	0.37	0.25	0.41	0.7	1.18	0.34	0.84	0.67	0.26	0.11	0.49	0.86	2.61
6	0.5	0.16	0.51	0.68	6.15	2.07	4.09	1.19	0.64	0.86	1.33	2.58	13.19

Table 22. Newly assigned player positions' statistics by Hierarchical Clustering method

Cluster 1 consists of rotational big men. These players are put on the court to play defense, grab rebounds, and hustle for loose balls. Players like Nic Claxton, Jalen Duren, Tyler Zeller, and Marreese Speights are all in this cluster. Cluster 2 consists of franchise players that are focal points of the teams. They usually play the most minutes, score the most points at an average of nearly 19 points, take the greatest number of shots, and are just statistically impressive in all categories that consequently make them stars. As shown in Table 23, Michael Jordan, Kevin Durant, Lebron James, Luka Dončić, Joel Embiid, and Kobe Bryant all belong to this cluster.

player	mpg	ppg	fg_percent	x3p_percent	ft_percent	trpg	apg	spg	bpg	season
Michael Jordan	37.922	28.926	0.4889	0.3508	0.8268	6.249	4.934	2.1348	0.663	10
Kevin Durant	36.758	27.224	0.4961	0.3834	0.8845	7.0634	4.2928	1.0867	1.1121	15
LeBron James	38.16	27.126	0.5046	0.3453	0.7343	7.4803	7.3535	1.5634	0.7617	20
Luka Dončić	33.881	26.6	0.4585	0.3344	0.7389	8.5667	7.9778	1.0667	0.4222	5
Joel Embiid	31.278	26.042	0.4903	0.337	0.8096	11.344	3.2934	0.8623	1.6437	7
Allen Iverson	40.698	25.805	0.4261	0.3126	0.7817	3.6081	6.0977	2.0799	0.1739	20

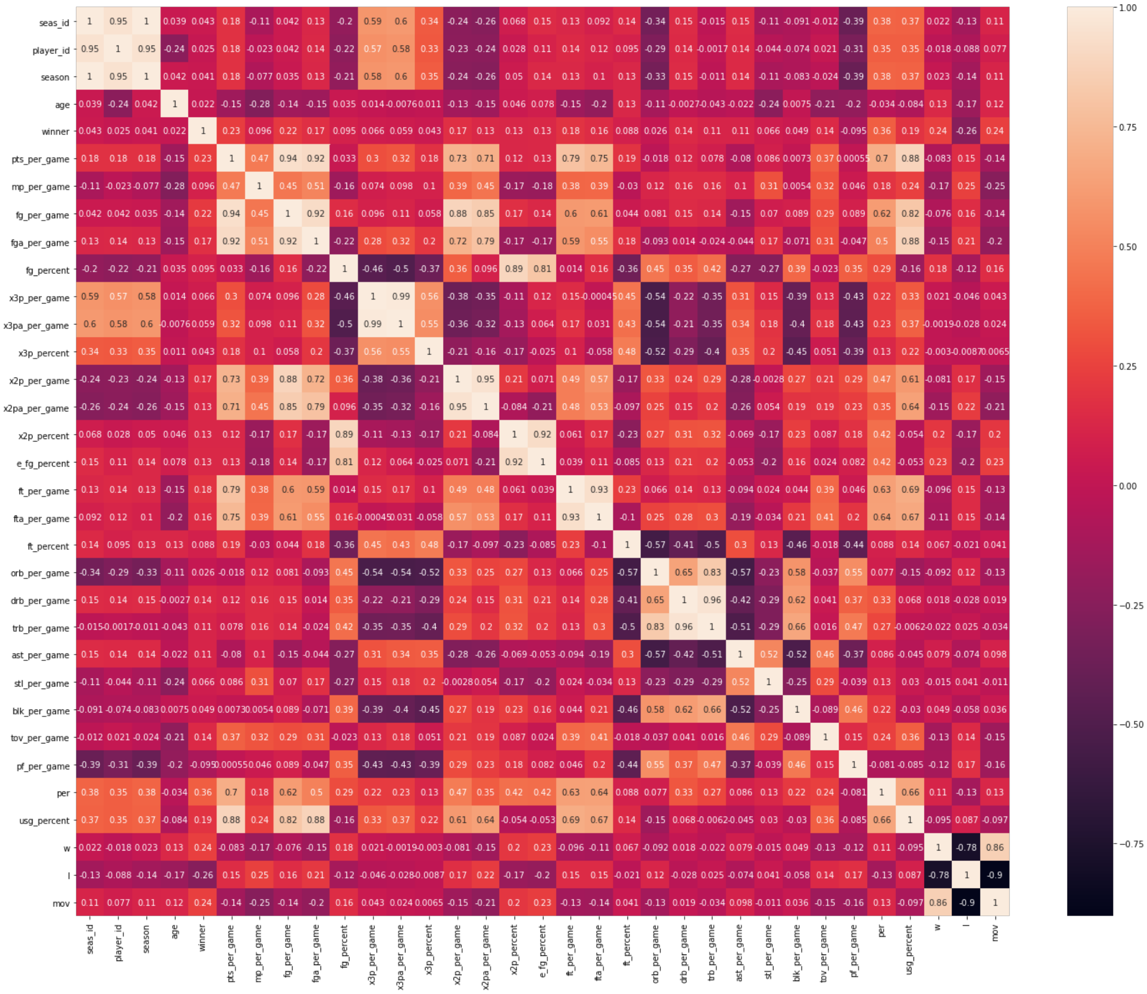
Table 23. Franchise Players

Cluster 3 consists of versatile wings that specialize in scoring and are great at three-point shots with an average of 34% three-point field goal percentage. Klay Thompson, Kawhi Leonard, Ray Allen, Glen Rice, Reggie Miller, and Jason Tatum all represent this cluster quite fittingly. These players also have longer careers compared to franchise players. Cluster 4 consists of versatile guards that are great rotation players. Despite having fewer minutes and points per game, these players shoot three-point shots as well as players in cluster 3. Distinct players like Danny Ainge, Tyrese Maxey, Jalen Brunson, and Cameron Johnson are all in this cluster. Cluster 5 consists of role players that don't have enough playing time and touches and have the shortest career at an average of 2.6 years. Lastly, cluster 6 consists of versatile big men that anchor the defense of the team. They have the longest career among all clusters at 13.19 years, block nearly as many shots as some franchise players do and have the highest field goal percentages at 50% as they often are responsible for baskets under or close the rim. Al Jefferson, Jermaine O'Neal, Rudy Gobert, Kenyon Martin, and Zydrunas Ilgauskas represent this group quite well.

Though the results do not agree with Alaggapan's claim, and Bianchi, Facchinetti, and Zuccolotto's findings, the model still produces distinct player categorizations with the statistics presented.

Can we predict league MVP using historical performance data?

The highest individual honor that an NBA player can have in his NBA career is the most valuable player (MVP) award. To explore whether this award can be determined more objectively through statistical analysis rather than having journalists and fans vote which is more prone to bias, several machine learning techniques are implemented. Before the models are trained, the dataset is first checked for multicollinearity. Table 24 suggests multiple features are indeed highly correlated.



	seas_id	player_id	season	age	winner	pts_per_game	mp_per_game	fg_per_game	fga_per_game	fg_percent	x3p_per_game	x3pa_per_game	x3p_percent	x2p_per_game	x2pa_per_game	x2p_percent	e_fg_percent	ft_per_game	fta_per_game	ft_percent	orb_per_game	drb_per_game	trb_per_game	ast_per_game	stl_per_game	blk_per_game	tov_per_game	pf_per_game	per	usg_percent	w	l	mov
seas_id	1	0.95	1	0.039	0.043	0.18	-0.11	0.042	0.13	-0.2	0.59	0.6	0.34	-0.24	-0.26	0.068	0.15	0.13	0.092	0.14	-0.34	0.15	-0.015	0.15	-0.11	-0.091	-0.012	-0.39	0.38	0.37	0.022	-0.13	0.11
player_id	0.95	1	0.95	-0.24	0.025	0.18	-0.023	0.042	0.14	-0.22	0.57	0.58	0.33	-0.23	-0.24	0.028	0.11	0.14	0.12	0.095	-0.29	0.14	-0.0017	0.14	-0.044	-0.074	0.021	-0.31	0.35	0.35	-0.018	-0.088	0.077
season	1	0.95	1	0.042	0.041	0.18	-0.077	0.035	0.13	-0.21	0.58	0.6	0.35	-0.24	-0.26	0.05	0.14	0.13	0.1	0.13	-0.33	0.15	-0.011	0.14	-0.11	-0.083	-0.024	-0.39	0.38	0.37	0.023	-0.14	0.11
age	0.039	-0.24	0.042	1	0.022	-0.15	-0.28	-0.14	-0.15	0.035	0.014	-0.0076	0.011	-0.13	-0.15	0.046	0.078	-0.15	-0.2	0.13	-0.11	-0.0027	-0.043	-0.022	-0.24	0.0075	-0.21	-0.2	-0.034	-0.084	0.13	-0.17	0.12
winner	0.043	0.025	0.041	0.022	1	0.23	0.096	0.22	0.17	0.095	0.066	0.059	0.043	0.17	0.13	0.13	0.13	0.18	0.16	0.088	0.026	0.14	0.11	0.11	0.066	0.049	0.14	-0.095	0.36	0.19	0.24	-0.26	0.24
pts_per_game	0.18	0.18	0.18	-0.15	0.23	1	0.47	0.94	0.92	0.033	0.3	0.32	0.18	0.73	0.71	0.12	0.13	0.79	0.75	0.19	-0.018	0.12	0.078	-0.08	0.086	0.0073	0.37	0.00055	0.7	0.88	-0.083	0.15	-0.14
mp_per_game	-0.11	-0.023	-0.077	-0.28	0.096	0.47	1	0.45	0.51	-0.16	0.074	0.098	0.1	0.39	0.45	-0.17	-0.18	0.38	0.39	-0.03	0.12	0.16	0.16	0.1	0.31	0.0054	0.32	0.046	0.18	0.24	-0.17	0.25	-0.25
fg_per_game	0.042	0.042	0.035	-0.14	0.22	0.94	0.45	1	0.92	0.16	0.096	0.11	0.058	0.88	0.85	0.17	0.14	0.6	0.61	0.044	0.081	0.15	0.14	-0.15	0.07	0.089	0.29	0.089	0.62	0.82	-0.076	0.16	-0.14
fga_per_game	0.13	0.14	0.13	-0.15	0.17	0.92	0.51	0.92	1	-0.22	0.28	0.32	0.2	0.72	0.79	-0.17	-0.17	0.59	0.55	0.18	-0.093	0.014	-0.024	-0.044	0.17	-0.071	0.31	-0.047	0.5	0.88	-0.15	0.21	-0.2
fg_percent	-0.2	-0.22	-0.21	0.035	0.095	0.033	-0.16	0.16	-0.22	1	-0.46	-0.5	-0.37	0.36	0.096	0.89	0.81	0.014	0.16	-0.36	0.45	0.35	0.42	-0.27	-0.27	0.39	-0.023	0.35	0.29	-0.16	0.18	-0.12	0.16
x3p_per_game	0.59	0.57	0.58	0.014	0.066	0.3	0.074	0.096	0.28	-0.46	1	0.99	0.56	-0.38	-0.35	-0.11	0.12	0.15	-0.00045	0.45	-0.54	-0.22	-0.35	0.31	0.15	-0.39	0.13	-0.43	0.22	0.33	0.021	-0.046	0.043
x3pa_per_game	0.6	0.58	0.6	-0.0076	0.059	0.32	0.098	0.11	0.32	-0.5	0.99	1	0.55	-0.36	-0.32	-0.13	0.064	0.17	0.031	0.43	-0.54	-0.21	-0.35	0.34	0.18	-0.4	0.18	-0.43	0.23	0.37	-0.0019	-0.028	0.024
x3p_percent	0.34	0.33	0.35	0.011	0.043	0.18	0.1	0.058	0.2	-0.37	0.56	0.55	1	-0.21	-0.16	-0.17	-0.025	0.1	-0.058	0.48	-0.52	-0.29	-0.4	0.35	0.2	-0.45	0.051	-0.39	0.13	0.22	-0.003	-0.0087	0.0065
x2p_per_game	-0.24	-0.23	-0.24	-0.13	0.17	0.73	0.39	0.88	0.72	0.36	-0.38	-0.36	-0.21	1	0.95	0.21	0.071	0.49	0.57	-0.17	0.33	0.24	0.29	-0.28	-0.0028	0.27	0.21	0.29	0.47	0.61	-0.081	0.17	-0.15
x2pa_per_game	-0.26	-0.24	-0.26	-0.15	0.13	0.71	0.45	0.85	0.79	0.096	-0.35	-0.32	-0.16	0.95	1	-0.084	-0.21	0.48	0.53	-0.097	0.25	0.15	0.2	-0.26	0.054	0.19	0.19	0.23	0.35	0.64	-0.15	0.22	-0.21
x2p_percent	0.068	0.028	0.05	0.046	0.13	0.12	-0.17	0.17	-0.17	0.89	-0.11	-0.13	-0.17	0.21	-0.084	1	0.92	0.061	0.17	-0.23	0.27	0.31	0.32	-0.069	-0.17	0.23	0.087	0.18	0.42	-0.054	0.2	-0.17	0.2
e_fg_percent	0.15	0.11	0.14	0.078	0.13	0.13	-0.18	0.14	-0.17	0.81	0.12	0.064	-0.025	0.071	-0.21	0.92	1	0.039	0.11	-0.085	0.13	0.21	0.2	-0.053	-0.2	0.16	0.024	0.082	0.42	-0.053	0.23	-0.2	0.23
ft_per_game	0.13	0.14	0.13	-0.15	0.18	0.79	0.38	0.6	0.59	0.014	0.15	0.17	0.1	0.49	0.48	0.061	0.039	1	0.93	0.23	0.066	0.14	0.13	-0.094	0.024	0.044	0.39	0.046	0.63	0.69	-0.096	0.15	-0.13
fta_per_game	0.092	0.12	0.1	-0.2	0.16	0.75	0.39	0.61	0.55	0.16	-0.00045	0.031	-0.058	0.57	0.53	0.17	0.11	0.93	1	-0.1	0.25	0.28	0.3	-0.19	-0.034	0.21	0.41	0.2	0.64	0.67	-0.11	0.15	-0.14
ft_percent	0.14	0.095	0.13	0.13	0.088	0.19	-0.03	0.044	0.18	-0.36	0.45	0.43	0.48	-0.17	-0.097	-0.23	-0.085	0.23	-0.1	1	-0.57	-0.41	-0.5	0.3	0.13	-0.46	-0.018	-0.44	0.088	0.14	0.067	-0.021	0.041
orb_per_game	-0.34	-0.29	-0.33	-0.11	0.026	-0.018	0.12	0.081	-0.093	0.45	-0.54	-0.54	-0.52	0.33	0.25	0.27	0.13	0.066	0.25	-0.57	1	0.65	0.83	-0.57	-0.23	0.58	-0.037	0.55	0.077	-0.15	-0.092	0.12	-0.13
drb_per_game	0.15	0.14	0.15	-0.0027	0.14	0.12	0.16	0.15	0.014	0.35	-0.22	-0.21	-0.29	0.24	0.15	0.31	0.21	0.14	0.28	-0.41	0.65	1	0.96	-0.42	-0.29	0.62	0.041	0.37	0.33	0.068	0.018	-0.028	0.019
trb_per_game	-0.015	-0.0017	-0.011	-0.043	0.11	0.078	0.16	0.14	-0.024	0.42	-0.35	-0.35	-0.4	0.29	0.2	0.32	0.2	0.13	0.3	-0.5	0.83	0.96	1	-0.51	-0.29	0.66	0.016	0.47	0.27	-0.0062	-0.022	0.025	-0.034
ast_per_game	0.15	0.14	0.14	-0.022	0.11	-0.08	0.1	-0.15	-0.044	-0.27	0.31	0.34	0.35	-0.28	-0.26	-0.069	-0.053	-0.094	-0.19	0.3	-0.57	-0.42	-0.51	1	0.52	-0.52	0.46	-0.37	0.086	-0.045	0.079	-0.074	0.098
stl_per_game	-0.11	-0.044	-0.11	-0.24	0.066	0.086	0.31	0.07	0.17	-0.27	0.15	0.18	0.2	-0.0028	0.054	-0.17	-0.2	0.024	-0.034	0.13	-0.23	-0.29	-0.29	0.52	1	-0.25	0.29	-0.039	0.13	0.03	-0.015	0.041	-0.011
blk_per_game	-0.091	-0.074	-0.083	0.0075	0.049	0.0073	0.0054	0.089	-0.071	0.39	-0.39	-0.4	-0.45	0.27	0.19	0.23	0.16	0.044	0.21	-0.46	0.58	0.62	0.66	-0.52	-0.25	1	-0.089	0.46	0.22	-0.03	0.049	-0.058	0.036
tov_per_game	-0.012	0.021	-0.024	-0.21	0.14	0.37	0.32	0.29	0.31	-0.023	0.13	0.18	0.051	0.21	0.19	0.087	0.024	0.39	0.41	-0.018	-0.037	0.041	0.016	0.46	0.29	-0.089	1	0.15	0.24	0.36	-0.13	0.14	-0.15
pf_per_game	-0.39	-0.31	-0.39	-0.2	-0.095	0.00055	0.046	0.089	-0.047	0.35	-0.43	-0.43	-0.39	0.29	0.23	0.18	0.082	0.046	0.2	-0.44	0.55	0.37	0.47	-0.37	-0.039	0.46	0.15	1	-0.081	-0.085	-0.12	0.17	-0.16
per	0.38	0.35	0.38	-0.034	0.36	0.7	0.18	0.62	0.5	0.29	0.22	0.23	0.13	0.47	0.35	0.42	0.42	0.63	0.64	0.088	0.077	0.33	0.27	0.086	0.13	0.22	0.24	-0.081	1	0.66	0.11	-0.13	0.13
usg_percent	0.37	0.35	0.37	-0.084	0.19	0.88	0.24	0.82	0.88	-0.16	0.33	0.37	0.22	0.61	0.64	-0.054	-0.053	0.69	0.67	0.14	-0.15	0.068	-0.0062	-0.045	0.03	-0.03	0.36	-0.085	0.66	1	-0.095	0.087	-0.097
w	0.022	-0.018	0.023	0.13	0.24	-0.083	-0.17	-0.076	-0.15	0.18	0.021	-0.0019	-0.003	-0.081	-0.15	0.2	0.23	-0.096	-0.11	0.067	-0.092	0.018	-0.022	0.079	-0.015	0.049	-0.13	-0.12	0.11	-0.095	1	-0.78	0.86
l	-0.13	-0.088	-0.14	-0.17	-0.26	0.15	0.25	0.16	0.21	-0.12	-0.046	-0.028	-0.0087	0.17	0.22	-0.17	-0.2	0.15	0.15	-0.021	0.12	-0.028	0.025	-0.074	0.041	-0.058	0.14	0.17	-0.13	0.087	-0.78	1	-0.9
mov	0.11	0.077	0.11	0.12	0.24	-0.14	-0.25	-0.14	-0.2	0.16	0.043	0.024	0.0065	-0.15	-0.21	0.2	0.23	-0.13	-0.14	0.041	-0.13	0.019	-0.034	0.098	-0.011	0.036	-0.15	-0.16	0.13	-0.097	0.86	-0.9	1

Table 24. Features multicollinearity heatmap

Firstly, the training datasets are imported into the algorithm to fit an optimal support vector machine model. As suggested earlier, three parameters in the regularization parameter (C), kernel type, and gamma values are tested. After looping through the pre-specified parameter ranges of the three designated parameters, an optimal SVM model is formed, and the model suggests that:

- The optimal regularization parameter (C) is 10
- The best kernel is the Radial Basis Function (RBF)
- The best gamma value is 0.1

To check whether the dataset is linearly separable, the intercept value is examined. As a rule of thumb, a value of the intercept that is close to zero, or less than 0.0001 would indicate that the data is reasonably linearly separable. Since the intercept value of this model is 0.24, it is assumed that the data is not linearly separable. In Tables 25 and 26, The SVM model's confusion matrix shows that it accurately identified 2 MVP winners from 2016 onward and correctly excluded 76 non-winners. The 2 MVP winners that are successfully predicted are Stephen Curry and James Harden in 2016 and 2018 respectively. Additionally, the model produces zero Type 1 error and 5 Type 2 errors where the model failed to pick 5 past MVPs.

Season	Award	Player	Age	Team	Winner	Predicted
2018	MVP	James Harden	28	Houston Rockets	1	1
2016	MVP	Stephen Curry	27	Golden States Warriors	1	1

Table 25. NBA MVP prediction using SVM model (Pre PCA)

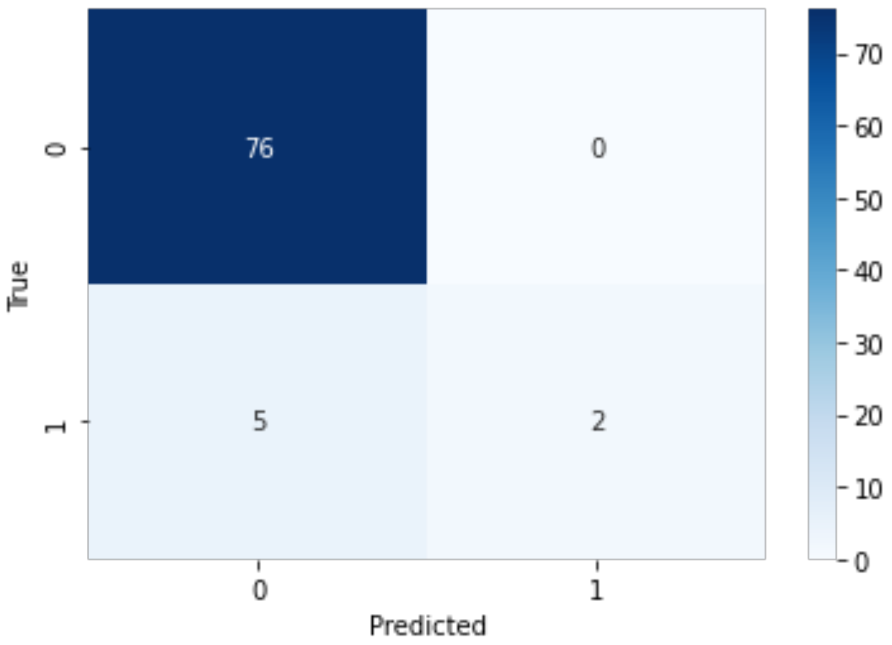


Table 26. Confusion matrix of SVM model (Pre PCA)

Next, the optimal Decision Tree model is built with the model suggesting:

- The maximum number of leaf nodes should be 10

In Tables 27 and 28, it is suggested that the Decision Tree model has higher accuracy than the SVM model as it successfully predicts one more MVP candidate in Nikola Jokić of the 2022 NBA season.

Season	**Award**	**Player**	**Age**	**Team**	**Winner**	**Predicted**
2022	MVP	Nikola Jokić	26	Denver Nuggets	1	1
2018	MVP	James Harden	28	Houston Rockets	1	1
2016	MVP	Stephen Curry	27	Golden States Warriors	1	1

Table 27. NBA MVP prediction with Decision Tree model (Pre PCA)

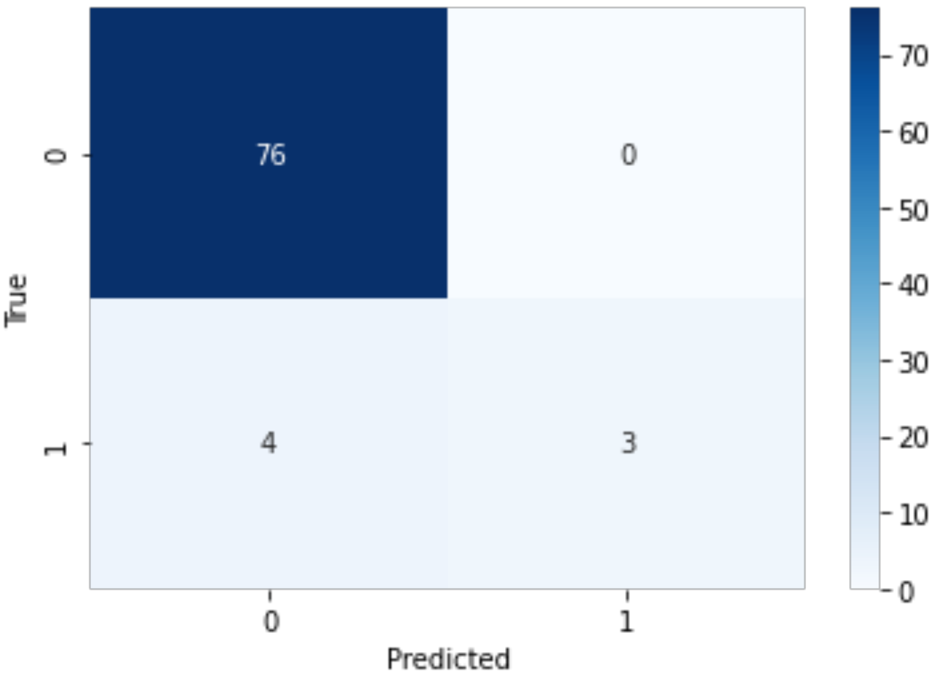


Table 28. Confusion matrix of Decision Tree model (Pre PCA)

Lastly, the optimal Random Forest model is built. The model suggests that:

- The number of features when looking for the best split in the forest is 3
- There should be 10 trees in the forest for optimal results

In Tables 29 and 30, the Random Forest model has the same accuracy as the Decision Tree model, though the model picks Giannis Antetokounmpo of the Milwaukie Bucks of the 2020 season instead of Nikola Jokić in 2022.

Season	**Award**	**Player**	**Age**	**Team**	**Winner**	**Predicted**
2020	MVP	Giannis Antetokounmpo	25	Milwaukie Bucks	1	1
2018	MVP	James Harden	28	Houston Rockets	1	1
2016	MVP	Stephen Curry	27	Golden States Warriors	1	1

Table 29. NBA MVP prediction with Random Forest model (Pre PCA)

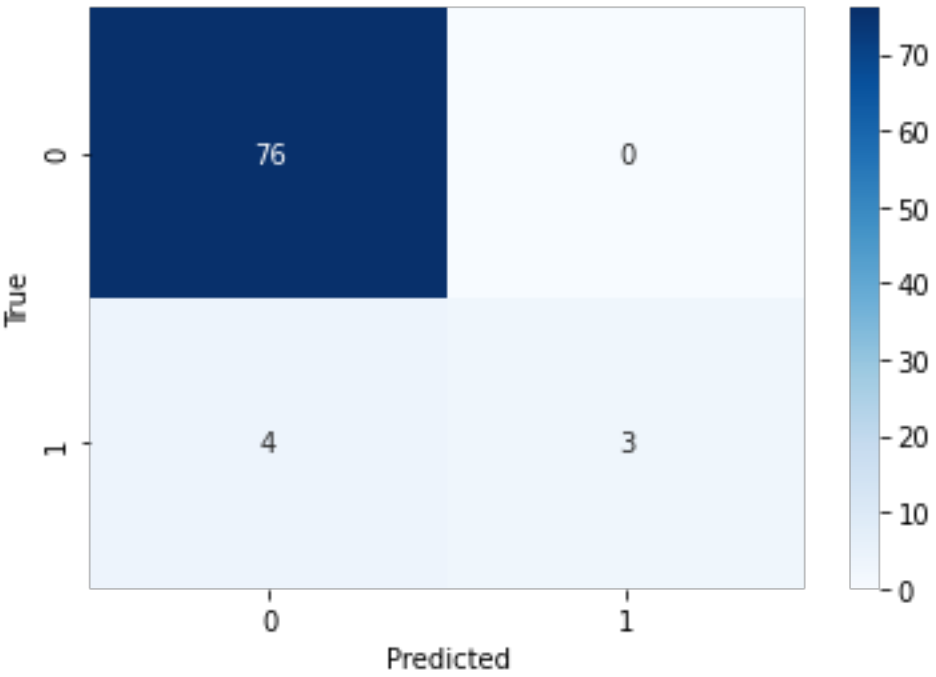


Table 30. Confusion matrix of Random Forest model (Pre PCA)

In Table 31, the accuracies, the performance metrics associated with the confusion matrix (True Positive Rate, Precision, Recall, False Negative Rate, and F1-Score), and the amount of error (MSE) are plotted. In short, the table indicates that Decision Tree and Random Forest are equally plausible at 95.2% accuracy, have higher true positive and false negative rates, and are less prone to error with the mean squared error of the two models at 0.048, compared to SVM's 0.06. Thus, at the first glance, both Decision Tree and Random Forest seem like good models to predict future MVP candidates.

Model	**Accuracy**	**Precision**	**Recall**	**FNR**	**F1-Score**	**MSE**
SVM	0.939759	1	0.285714	0.714286	0.444444	0.060241
Decision Tree	0.951807	1	0.428571	0.571429	0.6	0.048193
Random Forest	0.951807	1	0.428571	0.571429	0.6	0.048193

Table 31. Models Comparison (Pre PCA)

According to Table 32 and 33, 7 components produced by the PCA transformation should be preserved as the cumulative explained variance is 83.12%. On top of it, in Table 34, it is suggested that the PCA transformation has converted a collection of correlated variables from the original dataset into a set of uncorrelated variables.

Component	**Explained Variance**	**Cumulative Explained Variance**
1	0.2810	0.2810
2	0.2266	0.5076
3	0.0984	0.6061
4	0.0793	0.6854
5	0.0657	0.7511
6	0.0438	0.7949
7	0.0364	**0.8312**
8	0.0316	0.8629
9	0.0249	0.8877
10	0.0212	0.9090
11	0.0200	0.9289
12	0.0157	0.9446
13	0.0145	0.9591
14	0.0116	0.9706
15	0.0093	0.9799
16	0.0073	0.9873
17	0.0067	0.9940
18	0.0023	0.9963
19	0.0016	0.9979
20	0.0011	0.9989
21	0.0005	0.9994
22	0.0003	0.9997

Table 32. Explained Variance and Cumulative Explained Variance of the PCA Components

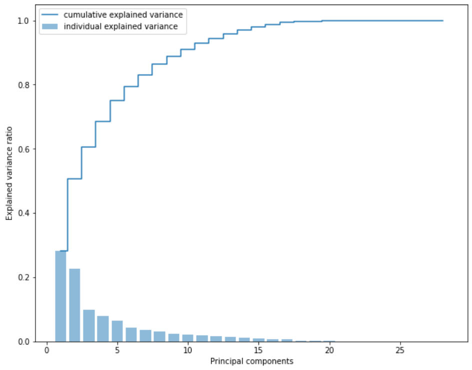


Table 33. Explained Variance by components

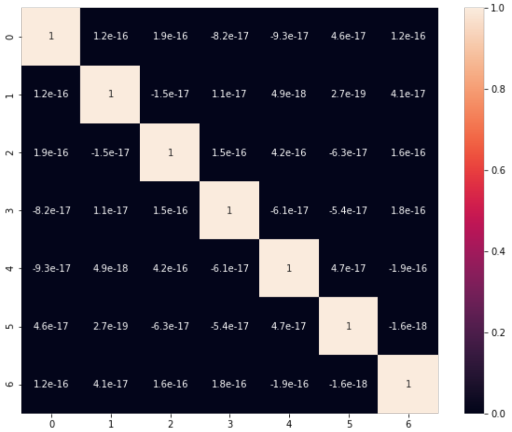


Table 34. Features multicollinearity heatmap after dimensionality transformation

Finally, each algorithm mentioned above is then trained again with the newly transformed dataset. The newly trained Support Vector Machine model suggests that:

- The optimal regularization parameter C should be 1
- The Best kernel is the linear kernel
- Gamma coefficient of 0.001 produces the best results.

On top of it, as shown in Table 35, the new SVM model successfully predicts four MVP candidates in Stephen Curry (2016), James Harden (2018), and Giannis Antetokounmpo (2019, 2020).

Season	Award	Player	Age	Team	Winner	Predicted
2020	MVP	Giannis Antetokounmpo	25	Milwaukie Bucks	1	1
2019	MVP	Giannis Antetokounmpo	24	Milwaukie Bucks	1	1
2018	MVP	James Harden	28	Houston Rockets	1	1
2016	MVP	Stephen Curry	27	Golden States Warriors	1	1

Table 35. NBA MVP prediction with Support Vector Machine (Post PCA)

In comparison, the optimal Decision Tree model with the transformed also suggests the maximum number of leaf nodes be 10, and as shown in Table 36, the model only successfully predicts two MVP candidates in Giannis Antetokounmpo in 2019 and 2020, one less than the model built before dimensionality transformation.

Season	Award	Player	Age	Team	Winner	Predicted
2020	MVP	Giannis Antetokounmpo	25	Milwaukie Bucks	1	1
2019	MVP	Giannis Antetokounmpo	24	Milwaukie Bucks	1	1

Table 36. NBA MVP prediction with Decision Tree (Post PCA)

Furthermore, the new Random Forest model indicates that:

- Four principal components should be included for the best split in the forest
- There should be 10 trees in the forest for optimal results.

Turns out, as shown in Table 37, the model predicts equal numbers of MVP winners, even though the winners are different that before the data transformation.

Season	Award	Player	Age	Team	Winner	Predicted
2020	MVP	Giannis Antetokounmpo	25	Milwaukie Bucks	1	1
2019	MVP	Giannis Antetokounmpo	24	Milwaukie Bucks	1	1
2018	MVP	James Harden	28	Houston Rockets	1	1

Table 37. NBA MVP prediction with Random Forest (Post PCA)

The new models' performances can also be further evaluated. Tables 38 and 39 suggested that the new SVM model seems to have the best performance of all three new models with an accuracy of 95.18% and recall (true positive rate) of 57.14%. The model has high precision and F1-score among all three models, meaning that the model correctly identifies a large proportion of positive cases that turn out to be correct, not to mention the lowest mean squared errors it produces. In contrast, the new Decision Tree approach generates poor outcomes, with a mere 91.56% accuracy and the lowest values for precision, recall, F1 score, and the highest rate of false negatives and mean squared errors.

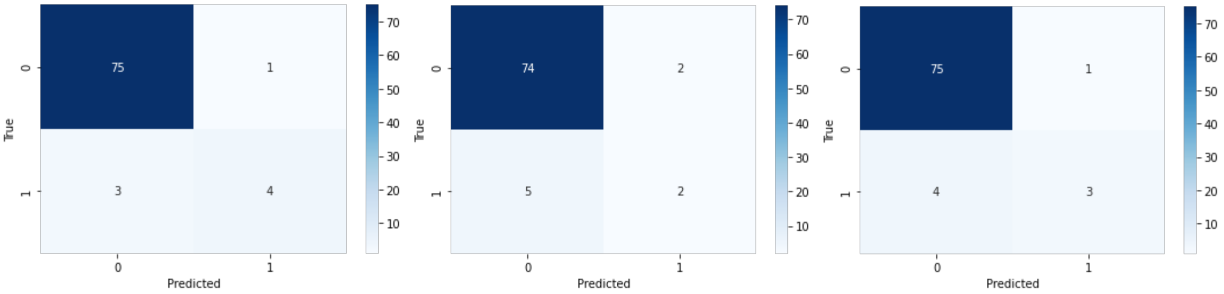


Table 38. Confusion Matrices of SVM, Decision Tree, and Random Forest models (Post PCA)

Model	Accuracy	Precision	Recall	FNR	F1-Score	MSE
SVM	0.951807	0.8	0.571429	0.428571	0.666667	0.048193
Decision Tree	0.915663	0.5	0.285714	0.714286	0.363636	0.084337
Random Forest	0.939759	0.75	0.428571	0.571429	0.545454	0.060241

Table 39. Models Comparison (Post PCA)

Lastly, to compare the three new models with the old models, the new SVM model seems to have the best performance out of all models at 95.2% and is the best model to be considered to predict future MVP winners. Though having a slightly lower precision rate at 80% compared to the old Decision Tree and Random Forest model that perform the best among the three old models, the new SVM model has the highest accuracy, recall, F1-score, and the lowest false negative rate and mean squared error.

VI. Conclusion

Looking back at the very argument raised by Stephen Jay Gould, to put things into perspective with the baseball analogy, the three-point shot field goal percentage is to the battling average, whereas the three-point shot field goal percentage standard deviation is to the batting average's standard deviation, or how dispersed the performances are from the mean. Gould suggested that there is an invisible right wall of biological limitations for humans in sports. In basketball terms, as seen in the analysis, despite having more players starting to shoot three-point shots, the field goal percentage plateaued at around 34% to 36% while the standard deviation continues to decrease, suggesting that NBA players are improving and catching up with the craft of three-point shots despite having "hitting that invisible wall".

With data science and sports analytics, the three-point shot is no longer just a "gimmick" that makes the game more exciting, but rather a legitimate tactic that teams can integrate deeply to gain advantages against the opposing teams. Just like the example of Houston Rockets introduced in the beginning, the implementation of the three-point shots only plays a part in a team's success and does not always guarantee wins and championships. However, through the lens of analytics, the values of three-point shots are displayed and highlighted, and players like Stephen Curry and Ryan Anderson are rewarded particularly well for their three-point shot craftsmanship. Nobody knows what other skill sets will prevail and accentuate, but with the evidence presented in this research, for the time being, three-point shots are here to stay.

On the other hand, not only the NBA continues to see changes and adaptations, the traditional player positionings no longer seem viable and can no longer represent what players do on the court. As suggested by Gould, basketball and other sports go through continuous and randomized changes. Players continue to be inspired and motivated by their predecessors, while the skill sets like three-point shootings continue to be reimagined and perfected. Teams on the other hand continue to evolve and be creative with the roster as players continue to get better and become more versatile. Young players that are aspired to be in the league continue to model their playing styles after the players that are currently in the league as the skill sets displayed are the ones that are apparently in demand by teams for the foreseeable future. With analytics and machine learning methodologies, it is not only capable of recognizing naturally occurring patterns within the players' statistics, but also identifying in-game features that are attributed to the new categorizations. And like many scholars and researchers suggested, not only is it possible to recategorize the majority of players into new groups with distinct characteristics, but teams can also utilize sports science more vividly to scout young prospects and to re-evaluate rosters formations.

Just like what Alaggapan, Bianchi, Facchinetti, and Zuccolotto's suggested in each of their findings, despite positions, the K-Means and Hierarchical clustering models presented in the analysis indicate that what hasn't changed at the core level for each team seem to be the presence of a franchise level or all-star level player that is being rostered throughout different eras. This type of player is characterized as having outstanding in-game statistics in categories such as points per game, minutes per game, field goal attempts and made, and other peripheral statistics. What changed in the player categorizations, are the types of players that the teams surround the franchise players with. Despite rule changes and shifting paradigms in what the league emphasizes at different points, whether it be the budget limitations or the complexity of different strategic directions, with sports analytics, teams are now able to examine and formulate a winning formula that best suits their circumstance, just like what the Oakland Athletics achieved in 2002 when a team filled with underdog players exceeded expectations and went on a historical winning streak.

Lastly, the MVP award symbolizes a player's greatness in all aspects of the game for that season. Nevertheless, even with players' performance statistics presented for comparison, the MVP debates among the sports media and individual fans are often polarizing, to say the least. Not only do media partners and fans look for compelling back stories behind each MVP candidate, but whether a player can elevate the team that he plays for is another determining factor, not to mention players with more fame or influence tend to be favored by the voters, making choosing an MVP winner a rather subjective and challenging.

In contrast, the methodology presented in the analysis simply offers a more objective alternative in choosing the MVP winners. With the historical performance statistics of all MVP candidates, the three predictive models that are trained with the data successfully exclude non-winners and pick a good number of the right candidates that won the awards from 2016 to 2022 with good accuracy. Though each with different model accuracies, the combination of using Principal Component Analysis and Support Vector Machine with the right parameter settings can predict the outcomes 95% of the time, making it a reliable model to predict future NBA MVP winners in the future.

In short, the increased use of sports science implies that teams look for qualities that can affect future results, whether it be winning games or championships as a team or developing and excelling in specific skills or aspects of the game as an individual player. Regardless, the methodologies presented in the analysis only serve as tools and guidelines to help teams and players to make sense of statistical information. Intangible attributes such as player's passion, interpersonal relationship skills, leadership, and many more are unquantifiable attributes. With sports science and analytics, teams can dictate strategic directions and identify problems more sophisticatedly in a data-driven manner, but ultimately it is up to NBA teams to make effort to understand their players on a personal level to effectively serve the interests of both the organizations and the players.

VII. References

1. Gould, S. J., 1986. Entropic homogeneity isn't why no one hits .400 anymore. Discover, August, pp. 60 – 66.
2. Gould, S. J., 1996. Full House - The Spread of Excellence from Plato to Darwin. New York, Harmony Books.
3. Chatterjee, S., & Lehmann, R., 1997. Evolution of team sports: A case study for National Basketball Association. Journal of Sport Behavior, 20(4), 412.
4. Moorefield, J., 2021. The Oakland Athletics use of sabermetrics and the rise of big data analytics in business. Honors Theses.
5. Neyer, R. (2017, August 20). Sabermetrics. https://www.britannica.com/sports/sabermetrics
6. Myers, M., 2013, November 18. The basketball analytics revolution should change your game. Ozy. Available from: https://www.ozy.com/news-and-politics/the-basketball-analytics-revolution-should-change-your-game/3381/ [Accessed 2 Jan 2023]
7. Buford, L. (2022, March 30). Miami Heat's Pat Riley was among the first fans of NBA analytics. Sports Illustrators. Available from https://www.si.com/nba/heat/miami-news/miami-heat-pat-riley-analytics [Accessed 2 Jan 2023]
8. Mills, J. (2015, June). Decision-making in the National Basketball Association: THE INTERACTION OF ADVANCED ANALYTICS AND TRADITIONAL EVALUATION METHODS.
9. Chung, J (2019, January). Explaining the trends of NBA strategy through the lens of human risk tolerance. (International Journal of Scientific & Engineering Research Volume 10, Issue 1)
10. Pelechrinis, K (2016, September 11). The Anatomy of the Three-Point Shot: Spatial Bias, Fractals and the Three-Point Line in the NBA
11. Moreno, E, Gil, D, & Vincent, F. (2022, December 8). Is the future of basketball being influenced by predictive data analysis? https://ssrn.com/abstract=4308292
12. Freitas, L. Shot distribution in the NBA: did we see when 3-point shots became popular?. Ger J Exerc Sport Res 51, 237–240 (2021). https://doi.org/10.1007/s12662-020-00690-7
13. Goldsberry, K. (2019, April 30). Sprawlball: A Visual Tour of the New Era of the NBA. Mariner Books; Illustrated edition.
14. Adams, L. (2016, July 28). Largest Free Agent Contracts of 2016. Hoops Rumor. Available from https://www.hoopsrumors.com/2016/07/largest-nba-free-agent-contracts-2016.html [Accessed 7 Jan 2023]
15. Bianchia, F., Facchinetti T., and Zuccolottob, P. (2017, November 15). Role revolution: towards a new meaning of positions in basketball. Electronic Journal of Applied Statistical Analysis Vol. 10, Issue 03
16. Chen, M. (2017). Predict NBA Regular Season MVP Winner. Proceedings of the International Conference on Industrial Engineering and Operations Management Bogota, Colombia, October 25-26, 2017
17. Albert AA, de Mingo López LF, Allbright K, Gómez Blas N. A Hybrid Machine Learning Model for Predicting USA NBA All-Stars. *Electronics*. 2022; 11(1):97. https://doi.org/10.3390/electronics11010097