FORECASTING MOST VALUABLE PLAYERS OF THE NATIONAL BASKETBALL ASSOCIATION

by

Jordan Malik McCorey

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Engineering Management

Charlotte

2021

Approved by:

_____
Dr. Tao Hong

_____
Dr. Linquan Bai

_____
Dr. Pu Wang

ABSTRACT

JORDAN MALIK MCCOREY. Forecasting Most Valuable Players of the National Basketball Association. (Under the direction of DR. TAO HONG)

This thesis aims at developing models that would accurately forecast the Most Valuable Player (MVP) of the National Basketball Association (NBA). R programming language was used in this study to implement different techniques, such as Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Linear Regression Models (LRM). NBA statistics were extracted from all of the past MVP recipients and the top five runner-up MVP candidates from the last ten seasons (2009-2019). The objective is to forecast the Point Total Ratio (PTR) for MVP during the regular season. Seven different underlying models were created and applied to the three techniques in order to produce potential outputs for the 2018-19 season. The best models were then selected and optimized to form the MVP forecasting algorithm, which was validated by predicting the MVP of the 2019-20 season. Ultimately, two underlying models were most robust under the LRM framework, which is considered the champion approach. As a result, two combination models were constructed based on the champion approach and proved to be most efficient. The two finalized combination models then served as the forecasting algorithm used to predict players PTR. Using this algorithm, one of the top players in PTR will win the MVP award for the regular season of the NBA. Hence, this proposed algorithm can be used post All-Star selections to determine the Most Valuable Player.

DEDICATION

To my Father and Grandfather

ACKNOWLEDGEMENTS

I would like to start by thanking God for putting me in this position in pursuing my graduate degree at the University of North Carolina at Charlotte. This experience has provided me a wealth of knowledge in the field of Engineering Management. Thank you to North Carolina A&T State University for providing me the foundation and basic skills of an engineer, as well as providing me essential experiences in my development. Thank you UNCC for prepping me throughout my graduate studies and providing me the opportunity to present this thesis capstone project. I would also like to thank my professor/thesis advisor, Dr. Tao Hong. When I initially joined UNC Charlotte's Mechanical Engineering Management program, learning about forecasting was not in the plans. When seeking for a capstone project, I wanted to do the thesis option to give myself enough time to provide a captivating and prominent piece of work. I came across the opportunity to base my project on sports forecasting from Dr. Hong. Certainly, I already had a passion for sports, especially basketball. I then connected with Dr. Hong, who advised me to take his forecasting class so I could learn about programming and different forecasting techniques. The class went extremely well, and I grasped on quickly, thus allowing me to apply that knowledge to my project. Ironically, Dr. Hong was also a basketball fan, and has been a strong pillar of support. Thank you to Dr. Linquan Bai and Dr. Pu Wang who served on my thesis committee. They also played a pivotal part in formulating my thesis, providing me with beneficial feedback and concepts to help improve my work. In summary, with the support from advisor and committee, I selected the thesis topic of forecasting the Most Valuable Player

of the regular season of the National Basketball Association and I can conclude that this was an exceptional project where I was able to apply my knowledge to my passion. Lastly, I would like to thank my family, girlfriend, and numerous friends who have endured this long process with me while always offering support and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

ABA	American Basketball Association
ANN	Artificial Neural Networks
APG	Assist Per Game
API	Aggregated Performance Indicator
AST	Assist
BLK	Blocks
BP	Back Propagation
BPG	Blocks Per Game
CLS	Constrained Least Squares
DEA	Data Envelopment Analysis
DM	Data Mining
DMUs	Decision Making Units
DPM	Defensive Plus Minus
DRB	Defensive Rebounds
DRtg	Defensive Rating
DWS	Defensive Win Shares
EFF	Efficiency
eFG	Effective Field Goal Percentage
ELO Rating	Team ELO
EWA	Estimated Wins Added
FG	Field Goals
FGA	Field Goals Attempted
FP	Fantasy Points
FT	Free Throws
FTA/FTr	Free Throws Attempted

GMs	General Managers
GmSc	Game Score
GP	Games Played
HACA	Hierarchical Agglomerative Cluster Analysis
IQ	Basketball Intelligence Quotient
KNN	K-Nearest Neighbor
LRM	Linear Regression Model
ML	Machine Learning
MP	Minutes Played
MPG	Minutes Per Game
MVP	Most Valuable Player
NBA	National Basketball Association
NBL	National Basketball League
NCAA	National Collegiate Athletic Association
NRtg	Net Rating
OLS	Ordinary Least Squares
OPM	Offensive Plus Minus
ORB	Offensive Rebounds
ORtg	Offensive Rating
OWS	Offensive Win Shares
PACE	Total number of possessions
PER	Player Efficiency Rating
PF	Player Fouls
PIE	Player Impact Estimate
PIPM	Player Impact Plus Minus
PIR	Performance Index Rating
PM	Plus Minus

PPG	Points Per Game
PPP	Points Per Possession
PRA	Points Rebounds Assists Percentage
PTR	Point Total Ratio
PTS	Points
REB	Rebounds
RPM	Real Plus Minus
SA	Simple Average
ScreenAssistPTS	Screen Assists to Points
SPG	Steals Per Game
STL	Steals
TL	Team Loss
TO	Turnover
TOV	Turnovers Percentage
TRB	Total Rebounds
TS	True Shooting Percentage
TW	Team Wins
USG	Usage
VA	Value Added
VORP	Value Over Replacement
WAR	Wins Above Replacement
WinsRPM	Real Plus Minus Wins
WS	Win Shares
3PAr	Three Point Attempts

INTRODUCTION

1.1 The Game of Basketball

Sport is defined as an activity involving physical exertion and skill in which an individual or team competes against another or others for entertainment. Sports have a significant influence on cultural experiences and have been documented for centuries. Some of the oldest sports in the world include wrestling, running/sprinting, gymnastics, polo, etc. Although sports are influenced by the culture and time period it emanates from, the impact of sports has been consistent since the beginning of time. With the ascendancy of its influence, sports have impacted economics, national unity, fan interest/pride, and have provided the public with role models and heroes.

Every pastime throughout history has produced exceptional athletes that the public today may consider to be the greatest of all time in their respected sport. The most effective way to compare athletes from different generations are through Sports Analytics [1]. Sport Analytics refers to the use of data and advanced statistics to measure performance and make informed decisions in order to gain a competitive sports advantage. The emergence of Sports Analytics begun August of 1971 from the Society for American Baseball Research and was initially considered a hobby [2]. The hobby has grown into a career for some, used in almost every professional sport; analyst collect data to increase revenue, improve player performance, team's quality of play, prevent injury and numerous other scenarios.

One of the world's most notable sports, is Basketball, otherwise known as hoops. This is a game played between two teams in which goals are scored by shooting a ball through a netted hoop on each end of a court; the team with the higher number of points wins the game. Dr. James Naismith is credited with creating the game of basketball in 1891 at the YMCA International Training School in Springfield, Massachusetts [3]. His intentions were to create an indoor activity that could be played during the winter months that would be fair for all players, and free of rough play. The game of basketball has matured into one of the most popular sports in the world where it is played in over one hundred countries. As the game has evolved, so has the athletes that play it. For this reason, the basketball world has seen legendary generational talents throughout its history. These individuals have revolutionized the game during their time and left an imprint both on and off the court. In the National Basketball Association (NBA), a player with a significant impact throughout a whole regular season can be recognized as the Most Valuable Player (MVP) in the entire league.

The evolution of basketball continues to excel to this day. The first intercollegiate basketball game was played between Hamline University and the Minnesota State School of Agriculture in Saint Paul, Minnesota on February 9th, 1895 [4], [5]. Just nine years later in 1904 the sport was played as a demonstration sport for the Summer Olympics. Professional basketball as the world knows it today was initially influenced by two leagues deemed the precursors of the NBA. The first precursor was the National Basketball League (NBL) which was in existence from 1938 to 1949. This league consisted of small Midwestern cities like Fort Wayne, Sheboygan and Akron. The NBL was one of the first leagues in America to provide opportunities for African American

players, and markedly, roughly 75% of today's NBA consist of players of African American descent [6]. The next forerunner league was the Basketball Association of America (BAA) which was in existence from 1946 to 1949. Unlike the NBL, the BAA established itself in bigger cities with larger major markets like Boston and New York. During the BAA's three years of existence, the league was in competition with the NBL for players and fans alike until they agreed to merge in 1949. Representatives from the two leagues met on August 3, 1949 to agree on joining together to form the National Basketball Association. The records and statistics of the BAA and NBL prior to the merger in 1949 are considered in official NBA history only if a player, coach or team participated in the newly formed NBA. There was also another professional league in existence from 1967 to 1976 that rivaled the NBA, it was the American Basketball Association (ABA). Due to the rivalry against the NBA, this league suffered financially which caused its termination and merge to the NBA. Though it ended, the ABA is also recognized with expanding the NBA in locations where the teams still exist to this day. The ABA also popularized the three point line; George Mikan, commissioner of the ABA, stated that the three pointer "would give the smaller player a chance to score and open up the defense to make the game more enjoyable for fans". Equally important, there are basketball leagues formed all across the world such as the Euro-League, Spain's Liga ACB, Turkish Basketball Super League, Russia's VTB United League, and many more. The impact of basketball is monumental, and the game has produced extraordinary players throughout its history.

1.2 Sports Analytics

Sports Analytics uses data collected during a game, throughout a season, and in the playoffs, to make evaluations and decisions. Sports analysts use the same basic methods as any other data analyst to manipulate and evaluate statistics collected throughout the season. This field provides an abundance of information accessible on multiple platforms for the consumers use. The market is continuously growing and is expected to reach almost $4B by 2022 [30]. The advancements in analyzing statistics have evolved sports themselves. Utilized by analyst, coaches, players, and fans alike, data analytics in the world of sports is an essential tool. For instance, sports analytics are used to evaluate the athletes, alternatively, it can serve as a way to engage fans, sell merchandise, price tickets, help teams improve and win, benefit the ecosystem, improve operational positions, and expand partnerships. It's important to explicate data variables in order to differentiate sports. For example, track measures a runner's speed to ensure they are performing at an adequate level to compete in their event, whereas soccer may measure player's stamina to determine how long they can continue to run and perform. There are varying metrics that sports analysts evaluate, but the data is ultimately used to advance the sporting experience.

Basketball is a popular sport that captures a plethora of data used to perform statistical analysis. Nearly every team in the NBA have data analyst on their staff who works with coaches and scouts to maximize their athlete's talents and to identify undervalued players [31]. There are so many variables involved with understanding the game of basketball, making it a challenging game to evaluate. First and foremost, NBA teams are limited to 15 players on a roster during the

regular season, meaning, basketball is a team sport. The composition of how teams are formed affects the team and individual player statistics through the season. In some cases, a player can have a strong enough influence in a game that it impacts his teammates as well as the opposing players. Additionally, anthropometrical measurements and fitness test results can be suitable data to predict players potential and skill on the court. Anthropometrical analytics can be influenced by a player's body size, strength, speed, agility, etc. Players of the same size and/or physical ability usually accumulate similar trends in their performance. Conversely, there are some players who have more skill than athleticism. Skill can be tied into a player's work ethic, repetition, basketball IQ, or just pure talent. Taking into consideration both the physical and mental parts of basketball can reveal the reason why players perform statistically bad or good through the regular season. In the early existence of the NBA, analyst kept track of player's basic statistical performance, things such as games played, points, assist, and rebounds. Basic statistics were the guiding principles of evaluating a player's game. Beginning in 2009, the league began using a video system to track movement of every player on the court, and the ball, 25 times a second [31]. The advancement in technology has given sports analyst an opportunity to accurately observe player's performance and record more reliable data. By the same token, as technology has advanced so has the statistics players are measured by. Basic statistics are still tracked today, but analyst also measure more advanced statistics that provide authentic assessments of players.

Analytics in basketball is used to study the game and its players so that teams can use that information to their advantage in hopes of having a successful season. The data that is collected

is aggregated and manipulated using varying techniques so that it can provide accurate assessments. Basketball statistics are available for anyone, starting from the collegiate level up to professional leagues. Anyone can conduct an analysis in basketball to try to support their cause, which depends on what data they're analyzing and the context that they're applying it too. Franchise owners and general managers benefit from analytics as well. They hire analysts to study data and trends of players on all levels of basketball, in hopes that they can bring in individuals that could improve their team. Owners and GMs use this information to fill their roster, surround their stars with quality role players, hire a coaching staff who can lead a team, exhilarate the fans to promote ticket and merchandise sells, and much more. Many of these responsibilities can be done with the use of analytics. Coaches use analytics to determine the skills of their own players and opposing players alike. They use the player's tendencies to set up schemes and gameplans in hopes of winning the game. Those who can effectively decipher players value through statistics can use that for their benefit. Players themselves also use analytics to measure their performance against other players. Fans also take data to make their own predictions and/or opinions of players, some may even use it to gamble. For instance, many people participate in March Madness for college basketball, some analyze and compare team statistics in order to forecast the results of bracket play for the NCAA. Forming estimations may never be guaranteed, but basketball analytics provide the best support in making forecast, whether that's predicting point spread, game winners, or player value.

1.3 Most Valuable Player

The NBA is considered the best professional basketball league in the world, because it has, and still produces the greatest players to ever play the sport. Players like Michael Jordan, LeBron James, Kareem Abdul-Jabbar, Kobe Bryant, Larry Bird, and Magic Johnson have had historical performances that have placed them on the Mount Rushmore of basketball legends. By the same token, there has been exceptional foreign players that have also played in the NBA such as Dirk Nowitzki, Hakeem Olajuwon, and Steve Nash. All the players listed above have been deemed as the Most Valuable Player at least once in their career. The MVP award in the NBA is an honor typically bestowed upon an individual as the most performing player in an entire regular season. This does not particularly mean that the MVP is the best player in the league, but the award is given to the individual who had the greatest impact individually and for their team. The award was introduced in the 1955-56 season. The winner of the award receives the Maurice Podoloff Trophy, which is named in honor of the first commissioner of the NBA.

An individual could never be in the conversation of being the greatest player of all time, a.k.a. GOAT, in basketball without winning the most valuable player accolade. MVP is the most prestigious individual award in basketball and achieving this solidifies that player in history for the entirety of their career and beyond. As a case point, the award usual goes to the most deserving player who had the most consistent and impactful play through a complete season. Yes, basketball is a team sport, and it has been proven that it takes a team to win a championship, which in most cases is the ultimate goal for all players. However, winning the MVP award for a

basketball player is a distinguished feeling. Achieving that honor validates the player's hard work and their worth to their team and the league. Winning this award provides motivation for that player to go chase a championship ring with the purpose of securing their legacy of being an ultimate champion. To date, there are only eight past MVPs who never won a championship; Charles Barkley, Karl Malone, Allen Iverson, Steve Nash, Derrick Rose, Russell Westbrook, James Harden, and Giannis Antetokounmpo. Granted, the last four names are still playing in the NBA today and have the opportunity to still win a title. This shows only 24% of all the past MVP recipients never won a championship. Winning it all not only supports the player's case for MVP but could justify their spot as a future hall of famer. Being granted the title of MVP means one of the four things for a player: 1) they are the best player in the league; 2) they own the narrative for that year; 3) they're the best player on the best team; or 4) they are the most valuable player on their team [28]. From the Merriam-Webster dictionary, the most valuable player is defined as the player who contributes the most to his team's success. Any player would be humbled to win the MVP award at any point in their career for this honor is awarded to an individual who is most deserving.

The voting mechanism for the MVP award of NBA has been evolving over the years. Up through the 1979-80 season, the players casted their votes for MVP. In the 1980-81 season, the voting power was redistributed to the media (sportswriters and broadcasters) throughout the United States and Canada. Starting in 2010, one ballot began being casted by fans through online voting. Voters rank five candidates, with corresponding points for each slot. First-place nets ten, second-place garners seven, third-place votes are worth five, fourth-place is three and fifth-place

earns one. This was a slight expansion of the system used during the league's early days, when voters ranked three players with respective point totals of five, three, and one [7]. The player with the most cumulative points is declared the MVP, in other words, the MVP is the player with the best point total ratio (PTR).

PTR is the dependent variable that will be measured in this thesis because it is the determining factor of who the MVP will be. This thesis will investigate why past MVP winners were bestowed the award. Different variables used to determine the MVP will be analyzed. An algorithm will be formed that will be able to accurately predict the PTR of players in the NBA. Obviously, a prediction is never 100% accurate for there is an infinite number of factors that influences the voter's ballot. Nonetheless, the algorithm will highlight the top three PTR values during the regular season, thus ensuring one of those individuals will be the MVP of that season.

1.4 Predicting Most Valuable Player

There have been 66 MVP awards presented to date. There are some commonalities between past MVP recipients, however, there are also intriguing differences that would put some recipients in a category of their own. Kareem Abdul-Jabbar (1975-76) is the only individual to win the award despite his team not making the playoffs. Karl Malone (1998-99) is the only recipient to win the award and not make an appearance in the All-Star game due to the event being canceled. The only two rookies to have won the MVP award are Wilt Chamberlain (1959-60) and Wes Unseld (1968-69). There are six instances of players who won MVP with a winning percentage under 60%; those athletes were Bob Petit (1955-56), Bob McAdoo (1974-75),

Kareem Abdul-Jabbar (1975-76), Moses Malone on two occasions (1978-79, 1981-82), and Russell Westbrook (2016-17). Stephen Curry (2015-16) is the only player to have won the MVP award by unanimous decision, whereas Shaquille O'Neal (1999-00) and LeBron James (2012-13) are the only two players to have fallen one vote shy of a unanimous selection. Some of these distinct occasions were investigated in this thesis to determine which statistical variables are significant when defining an MVP.

Primarily, players are voted on based off statistics that the athletes accumulate through the regular season. Players that excel in multiple statistical categories and can efficiently impact the sport throughout the entirety of a season, are notable candidates for MVP. One major statistical category that is glorified in basketball is scoring. There have been so many great scorers who have dominated in all scoring attributes from field goal percentage, shots attempted, free throw, and points. Another major statistical category is a player's efficiency. There are advanced statistics that are used to measure a player's efficiency such as PER, win share, and VORP. The MVP award is a multivariate honor that is seemingly difficult to determine. It is easy to recognize the best players during the regular season, but the challenge is to determine what factors separates the top tier players from the most valuable player. Of course, voters consider the statistics to sway their decision in who they will select, but there could be other determining components such as narrative or influence. How could Russell Westbrook win the MVP after averaging a triple double the year of 2016-17, then continue averaging a triple double the next two years and not even finish in the top three of MVP voting? How could James Harden lead the league in points from 2018 to 2020 and not win the award for MVP? Why is it that LeBron

James, the man considered by many to be the greatest player ever, not win MVP each and every year? The challenge of determining the MVP is that there is a plethora of variables that are compared amongst players. This thesis will explore those variables in order to accurately predict any players PTR, with the expectation of determining the most valuable player.

LITERATURE REVIEW

2.1 Sports Analytics

The first known use of Sports Analytics was captured in a book in 2003 titled Moneyball, by Michael Lewis. The book captures the quest of the Oakland Athletics general manager, Billy Beane, who realized using sabermetrics to obtain certain players that fit the preferred analytical criteria would bring success to the franchise [1]. This was the first platform where statisticians worked with individuals and team performance data (box score) to use for a competitive advantage [26]. In the novel, the Athletics utilized analytics to draft players who were able to get on base, compared to traditional measures like stolen bases or runs batted in. That insight provided a competitive edge in drafting great players overlooked by other teams [23]. This not only set a precedent for all the other baseball franchises, but this practice of using data analytics has expanded across professional sports as a whole.

Dr. Dave Schader, an experienced advanced developer/marketer for Teradata and knowledgeable sports analysis stated, “Sports analytics is the art and science of gathering data about athletes and teams for analysis to create insights that improve sports decisions, like deciding which players to recruit, how much to pay them, who to play, how to train them, how to keep them healthy, and when they should be traded or retired. For teams, it involves business decisions like ticket pricing, as well as roster decisions, analysis of each competitor’s strengths

and weaknesses, and many game-day decisions." [23].

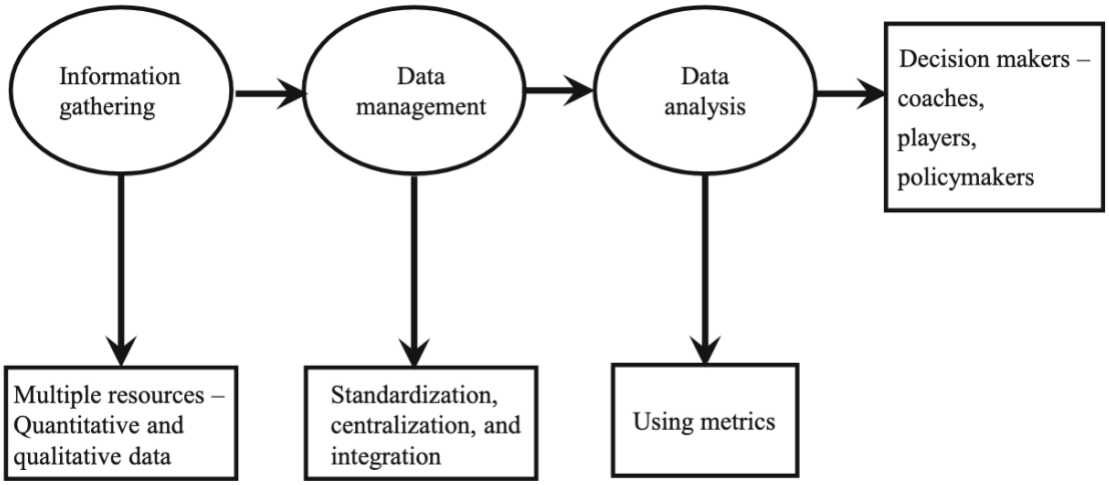

FIGURE 2. 1 Sport Analytic Framework

When dealing with large quantities of data available used to perform a study, the sports analytics framework, as seen in FIGURE 2.1, should be followed. The intention will be to aggregate the most effective data points that will provide a precise outcome for the study under consideration [26]. The practice of Sports Analytics is applied in all professional sports and is utilized so that teams can get ahead and achieve the ultimate goal of winning championships and developing superstar players. The growth in the field has become popular amongst organizations and fans alike for monetary purposes and legacy debates. Given the fact that Sports Analytics is continuing to grow in technology and methodologies, conclusions are drawn in a wide range of applications.

Sports Analysis are looked upon to deeply understand data that is collected so that it can help make decisions, improve the team's rosters and/or staff, evaluate player's performance, forecast future performance, minimize injuries, increase revenue and enhance the fan experience [8]. There are different methods that can be utilized to collect data in professional sports. The various methodologies of data collection include interviews/questionnaires, databases, biometrics, camera technology, real-time analytical technologies, data visualization and other measurable tools. The advances in technology have made data more accessible and easier to interpret. In baseball, PITCHf/x, HITf/x, and FIELDf/x video systems are used to capture and analyze pitching, hitting, and fielding, respectively. Sports VU, the video analytic system used in basketball is able to produce huge datasets of player' movements, ball touches, rebounds, and shot locations [26]. Having accurate and reliable data is essential when performing any type of sport analysis.

Notably, just acquiring a vast amount of data will not suffice when needing to perform analytical techniques. There are numerous software's that could be used in data analytics that would help organize and prepare the data so that it can be used to conduct a study. Many analysts use R, Python, Tableau, Power BI, Cloud and IOT technologies to assist in implementing different data science methods [9]. These applications are used as means to make formulating analytical models easy and convenient. To demonstrate, Dr. Schader provided an example of a simple analytic technique in his interview. His technique used decision trees with multiple factors to predict whether an offense in football would likely run or pass. The data used to formulate this decision tree would need to include a training set of offensive plays, the team's

personnel, and the offensive formations. Each level of the decision tree would have some type of if/or condition. For this example, percentage outcomes were provided indicating if the offense was running or passing [23]. This analytical technique could be utilized by teams in football when preparing for upcoming games so their players can recognize formations and assume what type of play might be ran. On the other hand, data-driven models can be more complex dependent upon the techniques used. Complexity can sometimes attribute to improving the accuracy of the model, resulting in more efficient outcomes.

In a science and sports article, a sample of 50 youth archers recruited from varying youth archery schemes completed a one end archery score test [27]. The test gathered standardize physical fitness and motor skill parameter measurements for data; specifically, these variables were hand grip, vertical jump, standing broad jump, static balance, upper muscle and core muscle strength, which resulted in an archery shooting score for each participant. These testing parameters were carried out and the scores of the archers were observed to be equally distributed. The complexity of this study was brought into play after the data was gathered and systematize. Hierarchical Agglomerative Cluster Analysis (HACA) was used to ascertain the grouping of the archers with regard to all the performance parameters measured in the study. HACA is an exploratory and unsupervised technique in which a hierarchy of clusters are formed starting from one observation. Subsequently, identical observations are merged into a single cluster as the hierarchy is built from the dataset. The number of clusters are displayed in a dendrogram where examinees were split into two categories based on similarities from the observations; the categories were high performance archers and low performance archers. To ascertain the

performance differences of the archers based on the parameters measured, the authors incorporated Mahalanobis' distance, t-statistic, and Cohen's d effect size analysis. The supervised learning for classification is carried out by means of machine learning methods, such as ANN (single hidden layer with ten neurons) and the KNN (fine Euclidean-based). Multilayered models, like the archer study is composed of multiple techniques, which convolutes the process of creating the model but also can provide favorable results. Nonetheless, whether complex or basic, there are a variety of ways to analyze any sport using qualitive and quantitative data.

As noted previously, there are many applications of data analytics in the realm of sports. Coaches use it to evaluate their players and their team, determine what areas need improvement and how to exploit their competitor's weaknesses. Scouts use analytics to assess players coming into the league to draft and/or sign, they also study other players that are in the league for a potential trade or sign. Players can use sport analytics to compare their own performance to others and highlight the areas they thrive in. General managers and owners utilize analytics to determine their players roster, coaching staff, revenue streams, fan attendance, etc. Lastly, fans also use sports analytics to compare players and support their debates with their peers. Specifically, for this thesis, the intent is to use data analytics in the field of professional basketball. The next section will highlight examples of data analytics used to evaluate players performance in the NBA.

2.2 Basketball Analytics

Basketball Analytics has been used throughout the history of the NBA for the purpose of capturing and recording player's performance. Franchises across the league have applied analytics to their day to day business in order to put together successful teams. Dependent upon a team's identity, they can use analytics to evaluate the athletes and bring in players who can address the needs of the team. Defensive minded teams usually find players who average high values in steals, blocks, or advanced defensive statistics. Some teams may look for a scorer that specialize in shooting threes or scoring in the paint. Analytics are used to estimate the value and potential of athletes and aid franchises into making decisions.

Sports analytics are not only used to evaluate players, but also to measure teams as a whole based upon their performance. There are 30 teams in today's NBA, which means there are a plethora of players, variables, and situations to decipher from when conducting a basketball analysis. There was a study that developed a Data Envelopment Analysis (DEA) method for evaluating the competitive performance of NBA teams with non-homogeneity on both input and output sides [19]. By opening the inner structure of decision making units (DMUs), they split them into types of homogeneous sub-units. Under an "Overall-Sub" framework, they proposed a conclusive evaluation model for obtaining the efficiencies of the collection of individuals. They then decomposed the overall efficiency by obtaining the sub-efficiencies of sub-units. By applying this forementioned method, the overall performance of NBA teams was obtained, and the empirical results validated that the efficiency decomposition is naturally unique without any

other additional conditions. Therefore, based on this framework, they could not only evaluate the efficiency of non-homogeneous DMUs, but also reveal the deep reasons for their inefficiency. Team's overall efficiencies were calculated and compared to determine the dominant teams versus the weak teams. Sub efficiency provided a more accurate analysis and aided in pinpointing the reason of any inefficiency. The final results can be used to suggest player trades, change team schemes, and determine player impact on a team.

Another study aimed to measure performance analytics used in the NBA by utilizing Machine Learning (ML) and Data Mining (DM) techniques to provide a qualitative and quantitative analysis for team owners, players, coaches, and technical staff to help them predict future situations [9]. As basketball continues to flourish, so does the advances in its data analytics; it's used to understand, analyze, forecast, and compare basketball statistics to minimize the possibility of uncertain events and increase forecasting accuracy. The formula presented in the paper was the Aggregated Performance Indicator (API), which combined several basketball metrics in order to measure players' production.

$$\mathrm{API} = \frac{\begin{bmatrix} \mathrm{RPM}(+/-) + \mathrm{PER} + \mathrm{PIE} + 4\mathrm{Factors} + \mathrm{NetRtg} + \mathrm{EFF} + \mathrm{PIR} + \mathrm{Tendex} \\ +\mathrm{BPM} + \mathrm{PIPM} + \mathrm{GmSc} + \mathrm{FP} + {}^{\mathrm{WS}}/_{48} + \mathrm{TeamELO} + \mathrm{eFG\%} + \mathrm{TS\%} \\ +\mathrm{VORP} + \mathrm{WinsRPM} + \mathrm{ScreenAssistsPTS} + \mathrm{WinsRPM} + \mathrm{ScreenAssistsPTS} \\ +\mathrm{PRA} + \mathrm{REB\%} + \mathrm{LooseBallsRecovered} + \mathrm{PPP} + \mathrm{ASTRatio} \end{bmatrix}}{30}$$

EQUATION 2. 1 Aggregated Performance Indicator

Out of the 30 variables used in the API formula, there were quite a few not originally considered for the algorithm of this thesis. Those statistics include WinsRPM, ScreenAssistPTS, eFG%, EFF, FP, GmSc, PIE, PIR, RPM, PIPM, PACE, REB%, PRA, WAR, Deflections, PPP, Loose Ball Recovered, AST/TOV, ELO Rating, Tendex, and Four Factors. The API formula was used to analyze 20 NBA players on the condition of participating in at least 30 games per season and at least 15 minutes per game to determine the players with the highest performance value in the 2017-18 and 2018-19 seasons. As a result, the formula was able to accurately evaluate the significance of the NBA players. Back Propagation (BP) Neural Networks was used and proved to be the most accurate forecasting method in comparison to the other methods mentioned in the paper. Neural Networks provided exceptional accuracy because it utilized the adjusted p-values to classify the outliers over/under performers with a threshold of 10% to avoid bias. As can be seen in the examples of this section, data collected in basketball can be used to in a variety of ways. We will see in the following section how it can be used to specifically predict the MVP of any given regular season.

2.3 Forecasting the MVP of the NBA

There are many key points that deem to be valuable when determining an MVP [9]. First, it must be known that sports data can be unpredictable because it's irregular and sparse. Sparse due to majority of the players not having a long career nor remaining in the same league and/or team. Equally, data is irregular because players throughout history have played in different generations, meaning the game has changed through time and it's difficult to compare. Considering these two

conditions, recent five to ten years of data should be emphasized when performing a forecast. It should also be noted that sports include two important variables, luck and skill. Luck is something random that an individual cannot predict. Whereas skill is the natural ability to highly excel in the sport. Both should be considered when creating a forecast algorithm in order to minimize inaccuracy. There are other attributes that should be in consideration as well when determining a player's value, such as players psyche, injury risk, clutch factor, physical condition, marketing, coaching, team makeup, team chemistry, etc. All things considered, the MVP award is a multivariate type of selection between players' character, performance analytics and team's worth in the league. Forementioned was the API formula, a sophisticated forecasting equation that could predict players performance based on qualitative and quantitative advanced analytics used in basketball. It provided insight on how to consider the varying perspectives that define an MVP and how to formulate a formula used for forecasting.

In another article, the authors created a novel MVP forecasting system using neural networks [10]. To form this system, they used 23 years of data from 1997 to 2019. They validated their forecasting method by predicting the MVPs for the 2009-10 season (LeBron James) and 2016-17 season (Russell Westbrook). The NBA has recorded data for decades so that analyst could evaluate players effectiveness on the court and predict outcomes. The authors used three different datasets. The first one contained totals of data the players recorded during the season. The second dataset consisted of advanced statistics that players accumulated during the regular season. Lastly, the third one was a mixed dataset containing attributes from each of the first two datasets. Each group had a total of 17 variables in each of the datasets, and the three mutual

statistics between all of them were players, games played, and minutes played throughout the entire season. There were a handful of metrics from this paper that were not originally considered in the algorithm of this thesis such as games started, 3PAr, FTr, ORB%, DRB%, AST%, STL%, BLK%, TOV%, USG%. The three datasets are used to train the Neural Network model to predict the results. The neural network model was used because it had better performance when using large amounts of data and can be used for supervised learning.

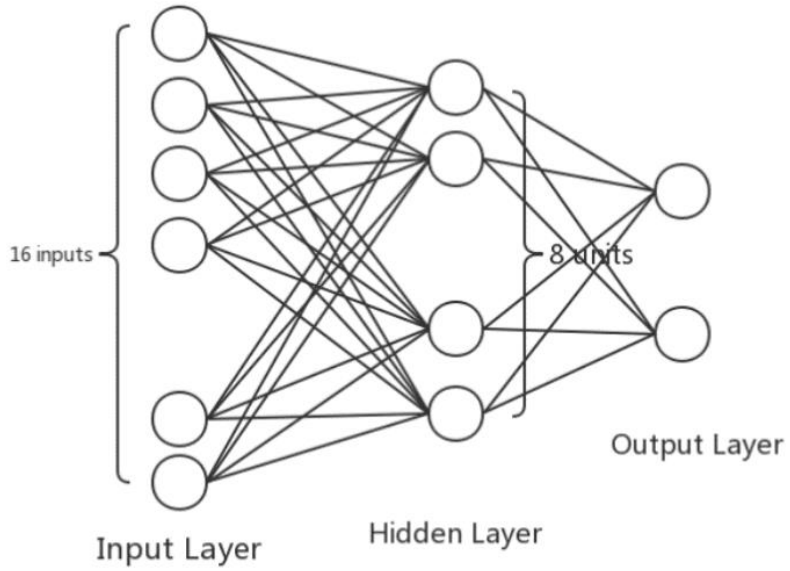

FIGURE 2. 2 Neural Network Layout

This specific neural network had three layers; first being the inputs of the neural network, second is the hidden layers representing a function that becomes like a neuron, and lastly, the output layer that presents the probability of becoming the MVP. The statistics from the datasets represents the input layer. The model was divided into three parts; training set, validation set, and

testing set. In their study, they used the validation set to evaluate and optimize the hyperparameter, called the learning rate in the neural network. To optimize their model, they used the mini-batch gradient descent algorithm. This computation works faster than other gradient descent algorithms and the jagged decline in the average cost function proves the mini-batch gradient is "kicking" the cost function out of local minimum values to reach better, perhaps even the best minimum. Lastly, the feed-forward back propagation technique was used in the neural network as well to process the input information through the nodes (neurons) until it was able to produce the output. The output was then compared to the expected value so that the error was calculated. The derivative of the error with respect to each weight provides the propagation which is then subtracted from the weighted value. Utilizing their models and data, they were able to compare the forecasted results for the 2009-10 and 2016-17 seasons. The charts presented in that work showed that the player who won the MVP of those years led in probability. The authors concluded that the mixed datasets had much better results than the other two datasets.

Ultimately, there is a wide spectrum of the challenges at which analytics can be employed when dealing with the game of basketball. Teams use analytics as a tool in so many areas as seen throughout this section. The theory is that the most substantial practice when performing basketball analytics is player evaluation. To have the capability of measuring a player's performance given their production, and then taking that data to predict their potential is a useful advantage for any team. This thesis will exploit this theory and use statistics to accurately forecast the value of players in hopes that it can determine the most valuable player of a regular season.

THEORETICAL BACKGROUND

3.1 Linear Regression Model (LRM)

Regression analysis is a statistical tool for investigating the relationship between variables. It is commonly used to predict future events and understand which factors cause an outcome. The credit of creating Regression Analysis is given to R.A. Fisher, a renowned statistician of the 20th century, who combined the work of Carl Friedrich Gauss (least squares) and Karl Pearson (Regression) to develop a fully realized theory of the properties of least squares estimation[15]. Regression analysis is thought of as one of the most simplistic and commonly used forecasting techniques to date. The representation is a linear equation that combines a set of input values (x) to solve for the predicted output (y). Each input value has a scale factor, also known as a coefficient (C). As an example, a simple regression equation is shown below.

$$y = C_0 + C_1 x_1 + C_2 x_2 \dots C_i x_i$$

EQUATION 3. 1 Simple Regression Formula

Regression is used to identify the strength that predictor variables have on the endogenous variable. Using this technique, an analyst has the ability to determine how much the dependent variable changes with a change to one or more independent variables. Regression analysis can also predict trends and future values for interpretation and estimations. This technique has continuously grown through time thus barring different sub-methods such as logistic regression, nonparametric regression, bayesian regression, regression that incorporates regularization, and

linear regression. Linear regression was selected as one of the forecasting techniques used to analyze the study of this thesis. The data used to interpret linear regression models must be structured to best utilize the model for its intended function. It is essential that data be transformed to make the relationship between the input and output variables linear. Linear regression works best with cleansed data, data in which does not contain noise (random fluctuations in the time series about its typical pattern). Removing noise grants better accuracy for the output variable and removes any possible outliers. It is also wise to consider calculating pairwise correlations for highly correlated input variables to remove its correlativity and cease the model from over-fitting. To increase reliability from the model, the input and output variables should have a Gaussian distribution. Gaussian distribution, also known as normal distribution, is a probability allotment that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. To perform this technique, some transform (log or BoxCox) can be used to make the distribution more Gaussian looking. One final method of data preparation for linear regression would be rescaling the input variables using standardization or normalization [29]. It must be emphasized that linear regression is a supervised learning model where predictions are formed from a chosen set of explanatory variables by effectively modeling a linear relationship. Some advantages include simple implementation, performance on linearly separable datasets, and overfitting can be reduced by regularization. On the other hand, linear regression is prone to underfitting and sensitive to outliers. This was the first technique of three used to perform the analysis of this study.

3.2 K-Nearest Neighbor (KNN)

K-Nearest Neighbor is a method for classifying objects based on the closest training examples in the feature space. This technique is considered to be a supervised machine learning algorithm where the function retains the entire training set during learning and assigns to each query a class represented by the majority label of its K-Nearest Neighbors in the training set [13], [14]. It is used to solve classification and regression problems, having either a discrete or real number value as its output. The distances between samples in the training set must be computed in order to determine unknown samples. The smallest value in distance corresponds to the sample in the training set closest to the unknown sample. Therefore, the unknown sample may be labeled based on the classification of the nearest neighbor. The most common way to find this distance is by the Euclidean distance formula, as shown below.

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_i - p_i)^2}$$

EQUATION 3. 2 Euclidean Distance Formula

KNN was introduced in 1951, where it was developed to be a non-parametric method for pattern classification so that it could perform discriminant analysis of probability densities that are unknown or difficult to determine. There have been many innovative contributions to KNN that has made the technique more efficient since its origin. Formal properties of KNN were established, rejection approaches incorporated, refinements with respect to the Bayes error rate, distance weighted approaches, soft computing methods and fuzzy methods were also added [13].

To emphasize, the performance of a KNN classifier is primarily determined by the choice of K as well as the distance metric applied. K is the specified number of examples selected, that are closest to the query. In order to select the right K for the provided data, the algorithm must run several times with different K values to compare its results. The K value that is best for the model is usually the one that reduces the number of errors while maintaining accurate predictions of new given data. There are some rules to consider when selecting the K value. If the K value holds the value of one, it should be expected that the prediction is less stable for there is not a sufficient evaluation of the surrounding values. Inversely, as the value of K increases, predictions become more accurate due to majority voting, averaging, etc. When errors start to appear, that is an indication that the K value is too high. KNN is an advantageous technique to use when forecasting because it is simplistic in regard to implementation, there's no need to build a model, one can tune several parameters, make additional assumptions, and the algorithm is versatile – used for classification and regression outputs. This technique's main disadvantage is that it gets significantly slower as the number of independent variables increase, which would not be an ideal method to use when a rapid prediction is needed. Other drawbacks of this technique include the trial and error method of finding the optimal K value and its poor performance at classifying data points within a boundary. Like LRM, the data used in the model must be cleaned and prepared so that the algorithm can read it properly. Data must be structured in a table format, for KNN assumes that all columns contain numerical data under the labeled rows. Missing values must be filled or removed as well before proceeding with this technique. KNN is often used in

simple recommendation systems, image recognition technology, and decision-making models; this was also one of the techniques chosen to conduct the analysis of this study.

3.3 Artificial Neural Networks (ANN)

Neural Networks were first modeled using electrical circuits by neurophysiologist Warren McCulloch and mathematician Walter Pitts in 1943 after writing a paper describing how neurons work [11]. Later, in 1949, Donald Hebb highlighted in his book, The Organization of Behavior, that neural pathways are strengthened each time they are used. Since then, the advances in Neural Network research have flourish profoundly. They evolved from just recognizing binary patterns, to adjusting weight values, to forming multilayered systems known as hybrid networks, to the ability of distributing pattern recognition errors throughout the network which is known as back propagation. In back propagation, the weights and thresholds are changed each time the model is ran such that the error gradually becomes smaller, thus signifying the accuracy of the network. This can be repeated hundreds of times until the error no longer changes. An Artificial neuron (perceptron) works the same as a biological neuron. The signals (input values) are numerical and multiplied by a weight so that the system can gather the needed information. Once all input values are calculated, the weighted sum is used to represent the total strength of the input signals and is then applied to a step function to determine its output. The outputs are then fed into other perceptron's dependent upon if the total strength exceeds the threshold. Below shows an equation displaying the scope of how ANN models work to produce the output.

$$Output_n = \sum weights_{n-1} * Input$$

EQUATION 3. 3 ANN General Rule Formula

Overall, the network is largely determined by the characteristics of the data. The model is constructed by a network of three layers of simple processing units connected by acyclic links. These models are trained to a significant amount of data used to classify new data based on what it thinks it's seeing. During the training period, the computed output is compared to the actual values. If values are the same, the ANN model is validated. On the other hand, if values vary, back propagation is activated to produce more accurate results. ANN is a flexible technique that can be applied to a wide range of time series forecasting problems with a high degree of accuracy [12]. This is a nonlinear data-driven self-adaptive method that can accurately predict future instances after learning the presented data. ANN models are universal approximators that can estimate any continuous function to any desired accuracy. The strengths of this technique include the effectiveness with nonlinear data and the non-requirement of prior knowledge of the process. Some weak points include limitations when attempting to forecast nonstationary data, its vulnerable to overfitting and that there is no guarantee the optimal solution will be predicted for all real forecasting problems. Neural networks have been applied to many problems in various fields, and it is deemed to be one of the more intricate methodologies in comparison to LRM and KNN. This study will compare all three techniques to determine the optimum method.

METHODOLOGY

4.1 Unanimous Votes

It is rare for a player to obtain all first place votes from the voters casting ballots for the MVP award. In fact, the only two players who fell short by one vote to a unanimous decision were Shaquille O'Neal (1999-00) and LeBron James (2012-13). Uniquely, there has been only one player to win the MVP award by unanimous decision, and that was Stephen Curry (2015-16). The question then becomes, what distinguished these players from all their competitors and fellow MVP recipients? This section will investigate how these three players earned their PTR and identify the variables that separated them from their peers.

TABLE 4. 1 Shaquille O'Neal 1999-00 Statistical Highlights

VARIABLE	PPG	PTS	FG	FG%	FTA	TRB	BLK	PER	WS	PM	VORP
VALUE	29.7	2344	956	0.574	824	13.6	239	30.6	18.6	9.3	9
League Ranking	1	1	1	1	1	2	3	1	1	1	1

TABLE 4. 2 Shaquille O'Neal 1999-00 Statistical Milestones

Variable	40+ Point Games	15+ Rebound Games	Double Doubles	Season High in Points
VALUE	9	33	63	61

Shaquille O'Neal won the Most Valuable Player award in the 1999-00 season after leading the Los Angeles Lakers to a 67-15 record, ranking first in the league. As seen from O'Neal statistical output in TABLES 4.1 and 4.2, it shows he was an unstoppable force and put up phenomenal numbers. The Lakers played through the big man, but he also had a young star in Kobe Bryant as his teammate who averaged 22.5 PPG that season. Having these two top caliber players made it extremely difficult for teams to guard both of them, which allowed Shaquille O'Neal to employ his dominance on his opponents. His influence led the Lakers to be the best rebounding team in the league and one of the best defensive teams as well. By the end of the season, the team averaged the lowest opponent FG% out of the 29 franchises during that time. The influence Shaq had on while he was on the court, as well as his personal achievements, led him to win MVP, one vote shy of a unanimous decision.

TABLE 4. 3 LeBron James 2012-13 Statistical Highlights

VARIABLE	PPG	PTS	FG	FG%	eFG%	TS%	AST	PER	WS	PM	VORP
VALUE	26.8	2036	765	0.565	0.603	0.64	551	31.6	19.3	11.7	9.9
League Ranking	4	3	1	5	2	3	8	1	1	1	1

TABLE 4. 4 LeBron James 2012-13 Statistical Milestones

Variable	30+ Point Games	10+ Rebound Games	10+ AST Games	Double Doubles	Triple Doubles	Season High in Points
VALUE	26	25	15	36	4	40

LeBron James won his fourth Most Valuable Player award in the 2012-13 season leading the Miami Heat to a 66-16 record, ranking first in the league. That season, LeBron James showcased his performance as shown through TABLES 4.3 and 4.4.The supporting cast included stars like Dwyane Wade, Chris Bosh, and Ray Allen who played significant roles for the team throughout the season. LeBron may go down in history as the best all-around player to ever play basketball, in Miami he learned how to control the pace of the game. He was efficient on both sides of the ball and made timely plays in every facet of the game, showcasing his basketball IQ. Furthermore, his impact led the team to have the best eFG% in the league and second best ORtg as a team. Ultimately, the way LeBron controlled the game throughout the season and his impact to his team granted him the MVP title, again, just one vote shy of a unanimous decision.

TABLE 4. 5 Stephen Curry 2015-16 Statistical Highlights

VARIABLE	PPG	PTS	FG	3P FG	3P FG%	eFG%	TS%	ASP	STL	PER	WS	PM	VORP
VALUE	30.1	2375	805	402	0.454	0.63	0.669	6.7	169	31.5	17.9	11.9	9.5
League Ranking	1	2	1	1	2	2	1	10	1	1	1	1	1

TABLE 4. 6 Stephen Curry 2015-16 Statistical Milestones

Variable	7+ 3P Games	40+ Point Games	10+ AST Games	3+ STL Games	Double Doubles	Triple Doubles	Season High in Points
VALUE	24	13	11	29	15	2	53

As previously stated, Stephen Curry is deemed the only MVP recipient to win the award by unanimous decision in the 2015-16 season where he led the Golden State Warriors to the best record in NBA history, 73-9. The team consisted of great players like Klay Thompson, Draymond Green and Andre Iguodala who gave Curry the opportunity to excel and shine. Additionally, Stephen Curry will go down in history as the best shooter to ever play the game. In fact, during the regular season of 2015-16, Curry set the NBA record at 402 three point field goals made. He evolved the game of basketball by his shooting efficiency, performing in a way never seen before in professional basketball. Stephen Curry broke both an individual record and team record all in one season which led him to be without a doubt the Most Valuable Player unanimously.

There are a few things that these three players had in common which won them the MVP. First, all these players led the league in four categories: PER, WS, PM, and VORP. These players also led their teams to at least 66 wins during the season. By the same token, they were able to excel individually while playing with other stars on their team. Ultimately, they all revolutionized the game in some sort. Shaquille O'Neal was an athletic big man who dominated the paint on both sides of the ball. LeBron James was an efficient point-forward who could control the game at his own pace while having an impact in every statistical category. Lastly, Stephen Curry was a marksman, being the most prolific shooter in the league's history. Using the analysis of these three players, some of the statistical categories were identified as significant.

4.2 Exploratory Analysis

There are two aspirations that players entering the NBA have, one is winning a championship at the highest level, and the second is winning the MVP award. These two accolades define a player's legacy. For any great player, legacy is important; the thought of solidifying their spot in the conversation of one of the greatest players to play the game is a remarkable accomplishment. The most valuable player award serves as affirmation for players hard work throughout a season. It certifies that their effort is not taken in vain, and their performance is to be recognized, to be forever cemented in the history books. Hence, one may ask, what variables in particular defines an MVP recipient? All things considered, there is a wide range of classifications that influences who deserves the MVP. This section will highlight the statistics captured in the NBA that are significant in determining value of a player.

Games Played (GP) is a vital statistic, and the first one that will be analyzed. This variable displays the total number of games a player has played in during the regular season. The purple vertical line in FIGURE 4.1 signifies the period of expansion where the number of games teams played during the regular season increased to 82 games. In the graph, there are three outliers, Bill Walton (1977-78) at 58, Karl Malone (1998-99) at 49, and LeBron James (2011-12) at 62. Bill Walton suffered a foot injury causing him to miss an extended period of time.

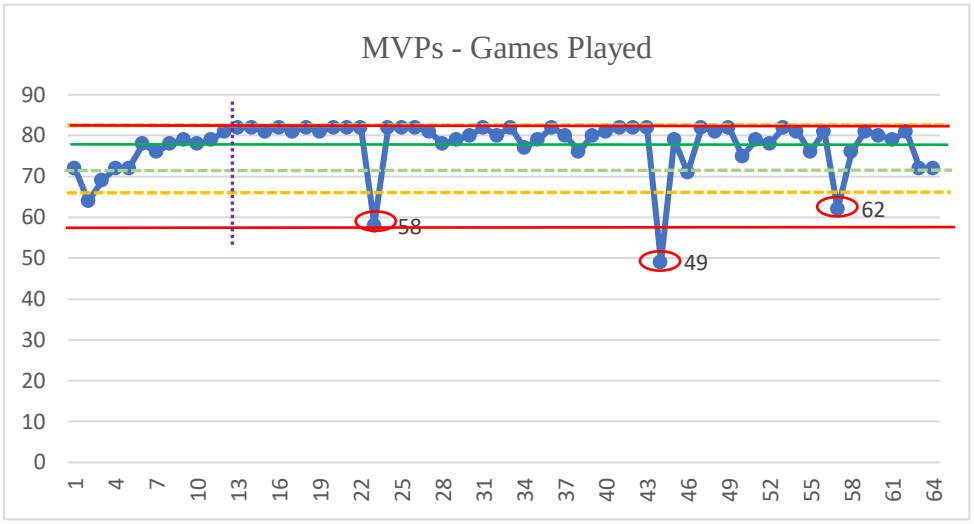

FIGURE 4. 1 Games Played by MVP Recipients

Conversely, both Malone and James experienced an NBA lockout season causing them to have a shorten season. Majority of all other MVP recipients played over 70 regular season games. Markedly, players who suffered a serious injury which caused them to miss significant amounts of time are highly unlikely to win the award. Minutes Played (MP) provides the number of minutes a player is on the court during the course of a season. Both GP and MP correlate with one another. Analyzing these statistics game by game is captured through Minutes Played Per Game (MPG), which provides the number of minutes a player averages per each game of the regular season. Minutes on the court provides players with the opportunity to perform and impact the game. Certainly, the most important variable for a franchise is winning, in this study both Team Wins (TW) and Team Losses (TL) were used in the analysis.

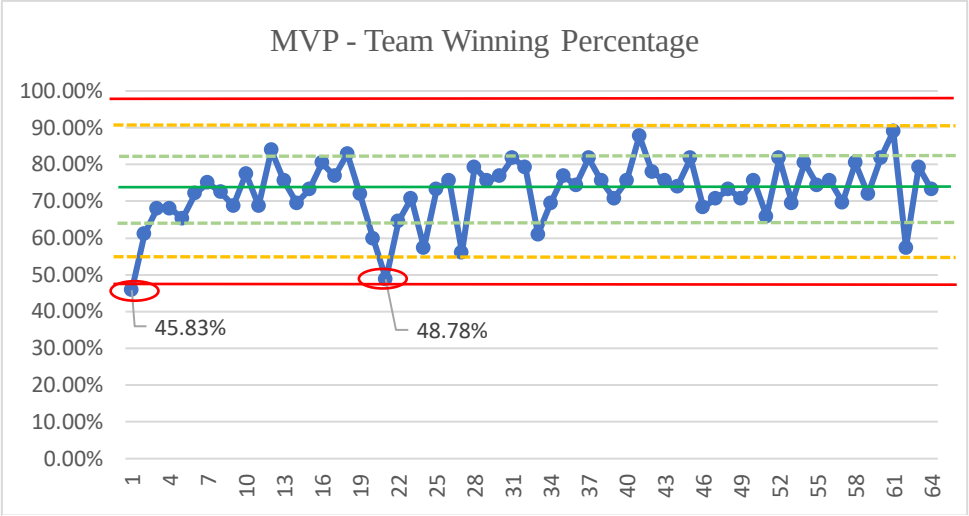

FIGURE 4. 2 Team Winning Percentage of MVP Recipients

The team winning percentage is a ratio of team wins over the total number of games played that season. There are two outliers in FIGURE 4.2, Bob Pettit (1955-56) at 45.83% and Kareem Abdul-Jabbar (1975-76) at 48.78%. Bob Petit was the first MVP in the history of the NBA. Though his team was ranked sixth in the league with a record of 33-39, he led the league in multiple major categories. Bob Pettit was the league leader in FG, FGA, FT, FTA, REB, PPG, and PER. His play did not bring success for his team, but he did shine as an individual, in fact, he may have been the best player in the league during that time. Interestingly, Bob had the lowest winning percentage of all time in comparison to all other MVPs. Through time, the emphasis on winning has grew exceedingly; it is highly unlikely that a player wins the MVP award today and not play for a winning team. Kareem Abdul-Jabbar won his fourth MVP title with the Los Angeles Lakers in the 1975-76 season after the team finished with a record of 40-42. Even

though the team was decent on offense, they were erroneous on defense, which seemingly caused them to lose more than half their games. Nonetheless, Kareem had an outstanding year, ranking first in the league in MP, REB, BLK, PER, WS, and PM. He was also runner up in the league in FG, FGA, and PTS. His team did not accomplish much that season, but his amazing performance throughout granted him MVP honors. Winning is the main purpose of the game. It is rare to have a player with a losing record earn the title of Most Valuable Player. A player must have a historic season to even be mentioned as an MVP nominee if his team has a poor team winning percentage.

Great players have an impact on the team's overall performance which can be shown through the teams rating. Team Offensive Rating (ORtg) is a metric that estimates how many points a team scores per 100 possessions; offensive productive efficiency; a measure used to evaluate team's offensive performance. Team Defensive Rating (DRtg) is just the opposite. Team Net Rating (NRtg) is the difference between the offensive and defensive rating. There are four outliers in FIGURE 4.3, which are Bob McAdoo (1974-75) at 0.1, Moses Malone (1981-82) at 0.0, Russell Westbrook (2016-17) at 0.8, and Michael Jordan (1995-96) at 13.4. Bob McAdoo's Buffalo Braves suffered losing key players in Gar Heard, Jim McMillian and Ernie DiGreorio which forced Bob McAdoo to take on more of the team's load. The team experienced some highs and lows throughout the season of 1974-75, ending with a record of 49-33. Bob McAdoo led the league in MP, FG, FTA, REB, PTS, WS and also ranked in the top percentile among BLK that season.

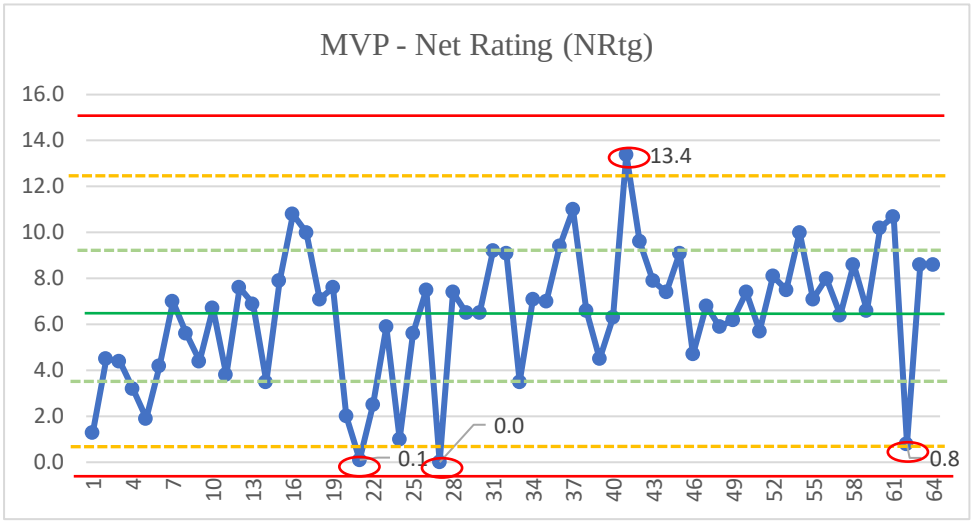

FIGURE 4. 3 Net Rating of MVP Recipients

Taking on the responsibility of carrying his team and elevating his game led McAdoo to the honor of winning MVP. Moses Malone's Houston Rockets did not have great team production during the season of 1981-82. The Rockets ranked 15th or higher out of the 23 teams in the league at the time in major statistical categories, particularly in FG, FG%, FT, FT%, AST, STL, and PTS. The Rockets also had the second slowest pace of any team that season. Moses however led the league in MP, FTA, REB, PER, WS, and OPM. The rockets finished the season with a 46-36 record. Russell Westbrook's Oklahoma City Thunder was a team that overestimated preseason predictions due to Westbrook's historic season, nonetheless, this team still had major insufficiencies. They ranked last in the league in three-point percentage, and 24th or higher in FT%, AST, TOV, and PF. Additionally, this team was weak defensively, especially in the paint, for many teams were able to exploit them for two-point field goals and free throws. Despite the

team's deficiency, Russell Westbrook remarkably averaged a triple double, setting a regular season record of 42. Westbrook also led the league in FG, FGA, PTS, PER, AST%, USG, PM, and VORP. The team ended the season with a 47-35 record. Michael Jordan's Chicago Bulls team is arguably the best team in basketball history, previously holding the best regular season record of 72-10 until surpassed by the 2015-16 Golden State Warriors. The only weak point for this team was free throws due to the lack of attempts the team put up. However, the team shot a decent free throw percentage of 74.6%. Jordan led the league in FG, FGA, PTS, USG, WS, PM, and VORP. He was the best player on the best team which easily won him the MVP title. Net rating is a significant statistic because it captures a team's performance. As seen above, a poor performing team can result in an outstanding individual seasonal performance. To become MVP, one will need to lead in major statistical categories throughout the league like Bob McAdoo and Moses Malone, or even break unfadable records like that of Russell Westbrook. It takes a special type of player to lead exceptional teams. The leader of a successful team will always be under consideration for MVP.

Other basic variables accumulated through the season on the defensive side of the ball include steals and blocks. Steals is when a player on the defensive end takes the ball or intercepts the ball from their opponent; in this study, Steals Per Game (SPG) was used in creation of the algorithm. Blocks are when a player on the defensive end deflects a field goal shot attempt and prevents their competitor from scoring; correspondingly, Blocks Per Game (BPG) was also used to form the algorithm. Both steals and blocks were not officially recorded until the 1973-74 season. Two detrimental variables in determining a player's value are turnovers and personal

fouls. Turnovers Per Game (TO) is when a player on the offensive end losses the possession of the ball. TO tracks how many turnovers a player makes on average per game. This statistic was not recorded until the 1977-78 season. A personal foul is a breach of the rules that involves illegal personal contact with an opponent. A player fouls out when they reach six personal fouls and is disqualified from participation in the remainder of the game. Personal Fouls Per Game (PF) tracks the number of personal fouls a player makes on average per game of the regular season. Moreover, Rebounds (REB) and Assist (AST) are considered essential statistics following scoring. An assist is when a player passes the ball to a teammate in a way that leads to a score by field goal, meaning that they were "assisting" in the basket. Assist Per Game (APG) provides the number of assist a player averages per each game of the regular season. Rebounds is when a player catches or retrieves the ball off a missed shot. Total Rebounds Per Game (TRB) captures the number of rebounds a player averages per each game of the regular season. Dependent upon what side of the ball the player is on when they retrieve the ball, rebounds can be separated between offensive (ORB) and defensive rebounds (DRB). ORB and DRB were not recorded until the 1973-74 season.

One of the most essential and glorious attributes in basketball is scoring. Scoring is recorded by points (PTS) where a player can score points by a two-pointer, three-pointer, or a free throw for one-point. One variable that was deeply analyzed was Points Per Game (PPG) which indicates the average expected number of points a player will provide each game.

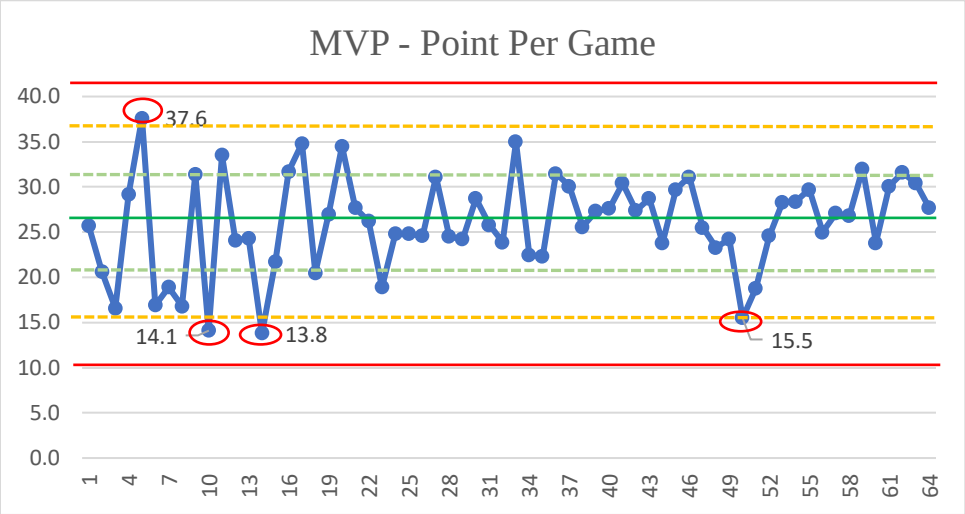

FIGURE 4. 4 PPG of MVP Recipient

FIGURE 4.4 shows four outliers, Wilt Chamberlain (1959-60) at 37.6, Bill Russell (1964-65) at 14.1, Wes Unseld (1968-69) at 13.8, and Steve Nash (2004-05) at 15.5. Wilt Chamberlin averaged 37.6 PPG in the 1964-65 season. Even as a rookie, Chamberlin showed his dominance standing 7-foot-1 nearly 275 pounds. He scored 50 or more points seven times that season including the playoffs. Additionally, he led the league in MP, FG, FGA, FTA, PTS, PER, and WS. His influence on offense was unlike any other athlete during that time period, thus granting him the honor of the Most Valuable Player. Bill Russell won the MVP in 1964-65 with only scoring an average of 14.1 PPG. The other contributing variables that won him the award was leading the league in REB and DWS. The team as a whole was dominant for they lead the league in FG, FGA, and REB; also came in top three in AST and PTS. The team had the best defensive rating and pace in the league. Bill Russell was the benefactor of the Boston Celtics culture and

the team's system in 1964-65. Wes Unseld averaged an all-time low in PPG out of all the MVP recipients. Despite his scoring, Unseld was able to win the award his Rookie year by the impact he provided to his team. The Baltimore Bullets had a 21 game improvement in comparison to the season before and ended the year in first place in the NBA with a 57-25 record. The team also finished first in FG, while making it in the top three of FGA, REB, PTS, DRtg, and pace. Unseld also finished second in REB and DWS for the year. His value may not have shown through his scoring, but the effect on the team was insurmountable. Steve Nash was crowned MVP the season of 2004-05 with scoring an average of 15.5 PPG. That year the suns had a league best 62 TW and 20 TL during the regular season. Steve Nash was the general for the Phoenix Suns as he led the entire league in AST averaging 11.5 per game. He was also ranked top ten in TS%, ORtg, OPM and OWS; and came in top 20 in PER and VORP. Certainly, his statistics are not glamorous compared to other MVP candidates that season, but his value to the team showed through the team's production. The Suns had the best ORtg that season, ranking first in FG, three-points made, REB, PTS, eFG% FG%, and pace. The purpose of basketball is for a team to score the most points, making scoring a significant attribute, but through the few cases seen above, there are other contributing variables that can prove to be just as valuable. History tells us that great scorers like Wilt Chamberlin are valuable assets to any team and could be a worthy candidate for MVP. However, a player who can make their teammates better and bring success to the whole team whether that's being a playmaker, a rebounder, and/or a defender will also make an individual a liable candidate for MVP as well.

There are advanced statistics that encompass scoring that were also considered in this study. A field goal in basketball is defined as a basket scored on any shot other than a free throw, worth two or three points depending on the distance of the attempt from the basket. Field Goal Percentage (FG%) is the ratio of field goals made to field goals attempted. In a like manner, Three Point Percentage (3P%) is the ratio of field goals made outside the three-point line to field goals attempted outside the three-point line. The three point shot was not relevant until 1979-80, thus being when the NBA incorporated the three-point line. An alternative way of scoring is free throws. Free throw shot(s) occur when a player commits a shooting foul, bonus foul, or technical foul on another player from the opposite team. Free Throw Percentage (FT%) is a ratio of free throw shots made to free throw shots attempted. True Shooting Percentage (TS%) measures a player's efficiency at shooting the ball. It is intended to more accurately calculate a player's shooting than FG%, FT%, and 3P% taken individually. Two- and three-point field goals and free throws are all considered in its calculation.

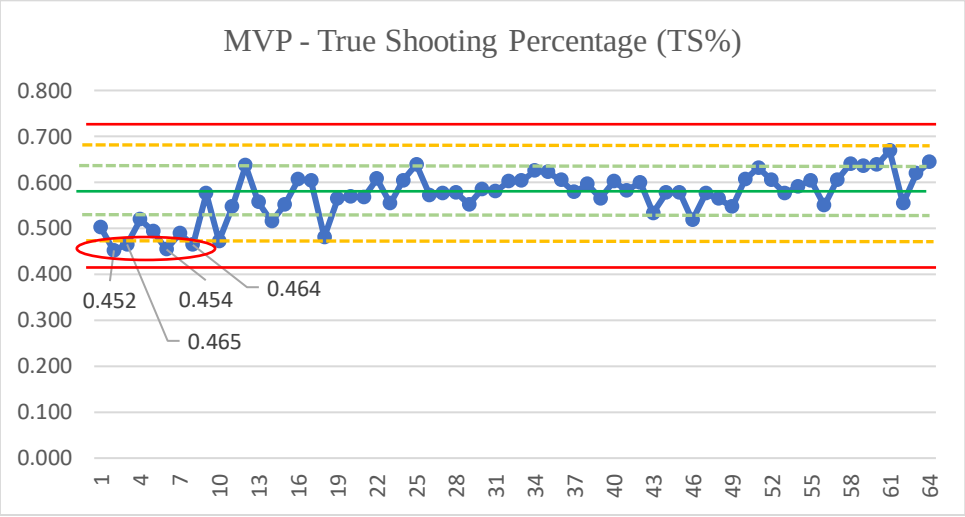

FIGURE 4. 5 True Shooting % of MVP Recipients

FIGURE 4.5 has four outliers in Bob Cousy (1956-57) at 0.452 and Bill Russell (1957-58, 1960-61 & 1962-63) at percentage values of 0.465, 0.454, 0.464. Bob Cousy was not the most prolific scorer during the 1956-57 season with the Boston Celtics. In fact, Bob Cousy only shot 37.5% from the field that season, 80.3% from the free throw, and had a PER of 21. Nonetheless, Boston had the best team record at 44-28, and was the best defensive team in the league. Bob Cousy ranked second in the league in DWS, meaning he was an essential part of the team's defense. Bob Cousy also ranked first in AST in the league, making him the key playmaker of a team with the most FG and PTS in the league. Bill Russell's field goal percentage for the three years cited above were 44.2%, 42.6%, and 43.2% accordingly. Though Bill Russell was not an efficient scorer, his teams during those years were dominating the league, finishing first in each of those seasons. The team exceled in defense, rebounding, and scoring as a team. Bill Russell was top in the league in DWS and REB through those years, which lead his team to success. These four outliers occurred early in the history of the NBA, but as of recently, true shooting percentage has been a significant variable. Bob Cousy and Bill Russell's Celtic teams were elite back in the 50s and 60s, and those two players were considered the leaders of those championship winning teams. This variable is a great representation of how characteristics that defines an MVP has evolved over time. Near the inception of the NBA, the MVP was seen as the most impactful player, whereas now the MVP is the most impactful player that can win and perform efficiently. The MVP may not be the best scorer in the league, but they do need to be an effective scorer and

key contributor on offense. As seen in FIGURE 4.5, the Most Valuable Player TS% has not been below 50% since the 1972-73 season.

The next advanced statistic examined in this study was Win Shares (WS), which is a metric that estimates the number of wins a player produces for his team. This attribute can be dissected by looking at WS between the offensive and defensive side of the ball.

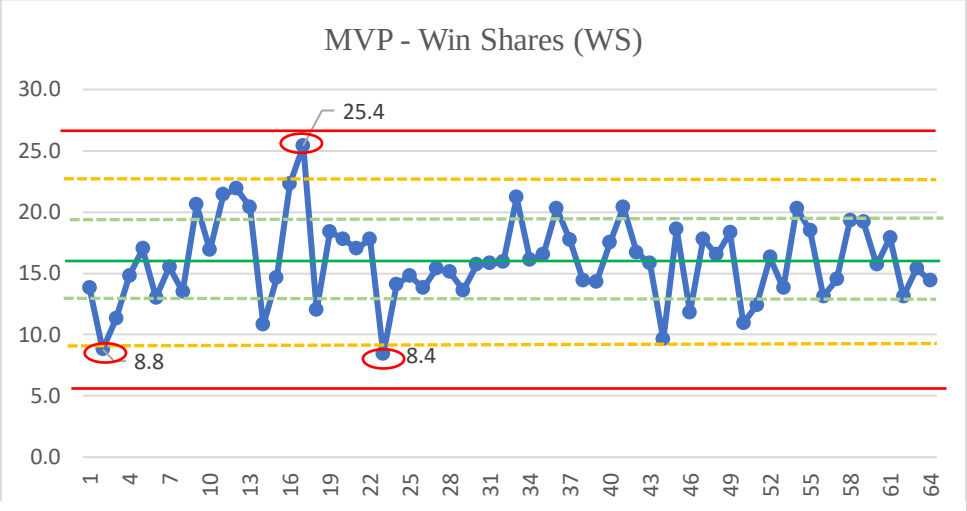

FIGURE 4. 6 Win Shares of MVP Recipients

FIGURE 4.6 shows three outliers in Bob Cousy (1956-57) at 8.8, Kareem Abdul-Jabbar (1971-72) at 25.4, and Bill Walton (1977-78) at 8.4. Bob Cousy was not the most efficient offensive player and that showed through his offense win shares at 4.0. The league high offensive win shares during that season were 11.6 by Neil Johnston. Nonetheless, Bob Cousy was a great defensive player having the second highest defensive win share at 4.7. Though he didn't win most of his offensive possessions, he was an essential defensive player for the best team in the league, and an imperative part to their success. Kareem Abdul-Jabbar was an influential force on

both offense and defense in the 1971-72 season. He led the league in FG, PTS, PER during that season as well as ranking top three in FTA, REB, FG% and MPG. His effort led the Milwaukee Bucks to have the second best record in the league at 63-19. Bill Walton had the lowest win share out of all the MVP recipients throughout history. Walton was a slightly above average player on offense but an outstanding defender, and led his Portland Trailblazers to a 58-24 record, league's best. Bill Walton showed his dominance through his defense and rebounding and was an intricate part to the team's success. Win Shares Per 48 mins (WS/48) shows how much the player contributes to a winning effort on a per game basis based on their per minute performance. Similar to win shares, this statistic shows a player contribution to winning every 48 minutes (per game). Win Shares is a favorable statistic to follow when determining an MVP candidate. More times than not, the most valuable player is winning their matchup in which contributes to team wins. Certainly, the higher the WS value is, the more well-rounded the individual player usually is. This is not a statistic required for a player to win the MVP, but is a great indicator given the past recipients.

The next major advanced statistic used in this study was Plus/Minus (PM) which reflects how the team performs by the points scored per 100 possessions while that player is on the court. This

statistic can be evaluated from an offensive aspect, as well as defensive. Furthermore, this was not an official recorded statistic until the 1973-74 season.

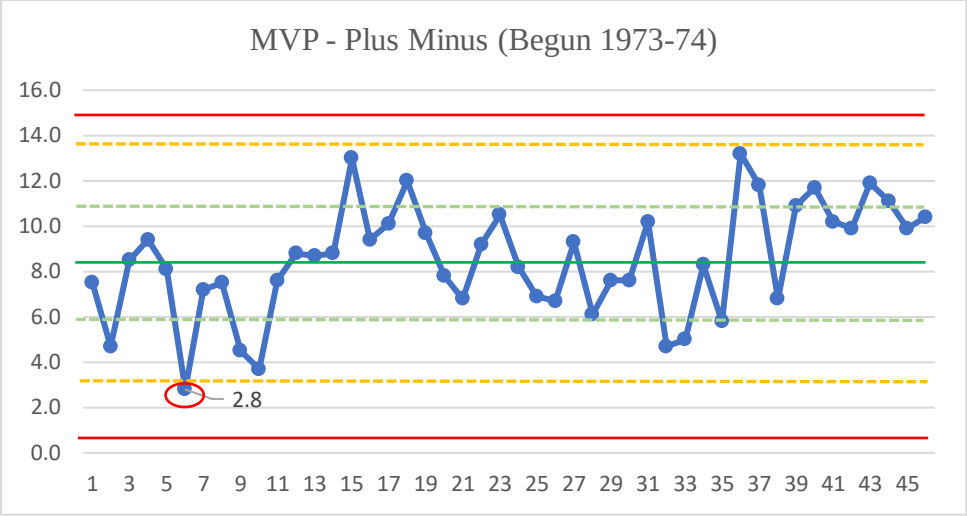

FIGURE 4. 7 Plus Minus of MVP Recipients

The only outlier for PM was Moses Malone (1978-79) at 2.8. Even though Moses Malone performed exceptionally in WS, PER, and TS%, his PM was surprisingly low. This was due to him being a defensive liability with a DPM of -1.7. He was a threat on offense; being a big body that's able to grasp rebounds and score in the paint. His lack of ability came when guarding his opponents, which made him a target each possession. This statistic acknowledges if a team has a positive or negative outcome while a player is on the court. It is expected for good things to happen when the MVP is on the court, for this reason, MVPs usually have higher values in PM.

Value Over Replacement Player (VORP) measures each player's overall contribution to the team versus what a theoretical "replacement player" would provide. The "replacement player" is

defined as a player on minimum salary or not a normal member of a team's rotation. This statistic was not officially recorded until the 1973-74 season.

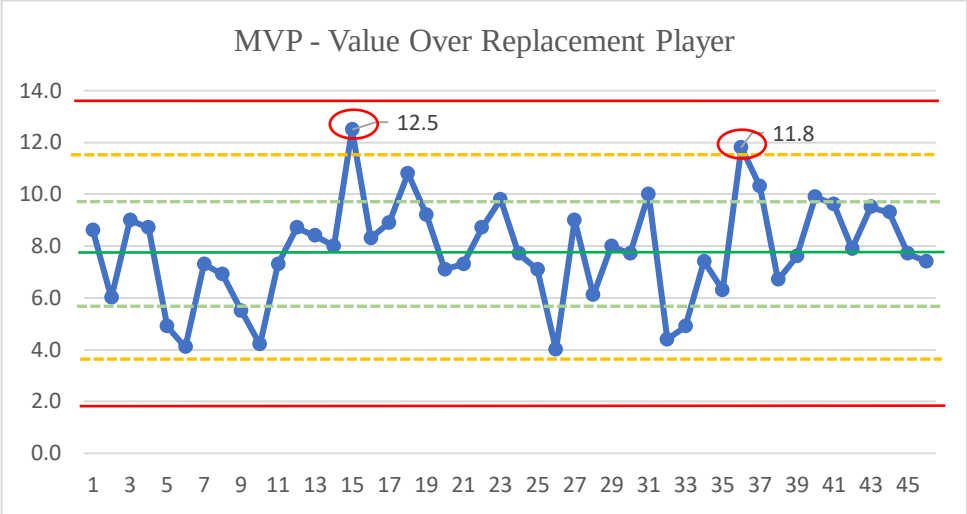

FIGURE 4. 8 VORP of MVP Recipient

The two outliers shown in FIGURE 4.8 are Michael Jordan (1987-88) at 12.5 and LeBron James (2008-09) at 11.8. Michael Jordan led the league in MP, FG, FGA, FT, STL, PTS, PER, WS, PM, and VORP. He was dominant both on offense and defense, while leading his team to a record of 50-32. His impact while on the court was immense and could not be replaced by an average player. LeBron James led the league in FT, PER, WS, PM, and VORP. He also came in the top three in FG, PTS, and USG for the year. LeBron led the Cavs to a 66-16 record, best in the league. He was the motor of that team and provided production in every way. These are arguably the two best players to ever play the game and they have the highest VORP throughout all the MVPs in history. This statistic emphasizes the importance of an individual player, for

valuable players cannot be replaced. This also leads into the next advanced statistic included in the study, Value Added (VA). VA is the estimated number of additional points a player contributes to a team's season, compared to the number of points a replacement player would contribute.

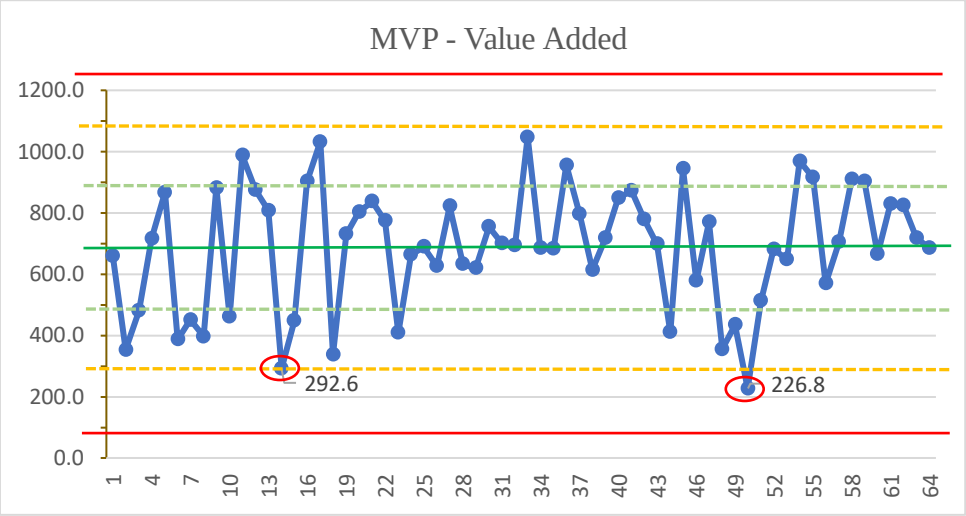

FIGURE 4. 9 Value Added of MVP Recipients

FIGURE 4.9 shows the two outliers, Wes Unseld (1968-69) at 292.6 and Steve Nash (2004-05) at 226.8. Wes Unseld was not a huge contributor on offense as stated before. He averaged about 6.6 more PPG and 1.7 more AST than the replacement player in his position. He averaged high MPG and was a strong rebounder which gave the team more point opportunities. Steve Nash was not a prolific scorer through his career, and in the 2004-05 season specifically, he only averaged 15.5 PPG. He was an excellent playmaker and created points for his teammates, averaging a league high 11.5 APG. However, Steve Nash lost points in TO at 3.3 and was a defensive liability with a DPM of -0.8. His replacement player was Leandro Barbosa who averaged 7 PPG,

2 APG, and 1.4 TO with a similar defensive ability as Steve Nash. Though he was nowhere the playmaker Steve Nash was, he could score consistently when given the minutes. For high volume scorers, this statistic tends to be high since they contribute a large portion of the team points each game. Contribution of points contributes to wins, and for this reason, Estimated Wins Added (EWA) correlates with VA. EWA estimates the number of wins a player contributes to a team's season total compared to what a "replacement player" would contribute.

Another key statistic is Usage Rate (USG) which measures the percentage of team plays used by a player while on the floor is. This variable was not officially recorded until the 1977-78 season. FIGURE 4.10 shows two outliers, Steve Nash (2004-05) at 20.5 and Russell Westbrook (2016-17) at 41.7. Steve Nash had the lowest usage rate of all the MVPs in history due to the type of player he was and how his team was composed. He is considered to be a traditional point guard, one that facilitates and makes plays for his teammates more than look for his own shot.

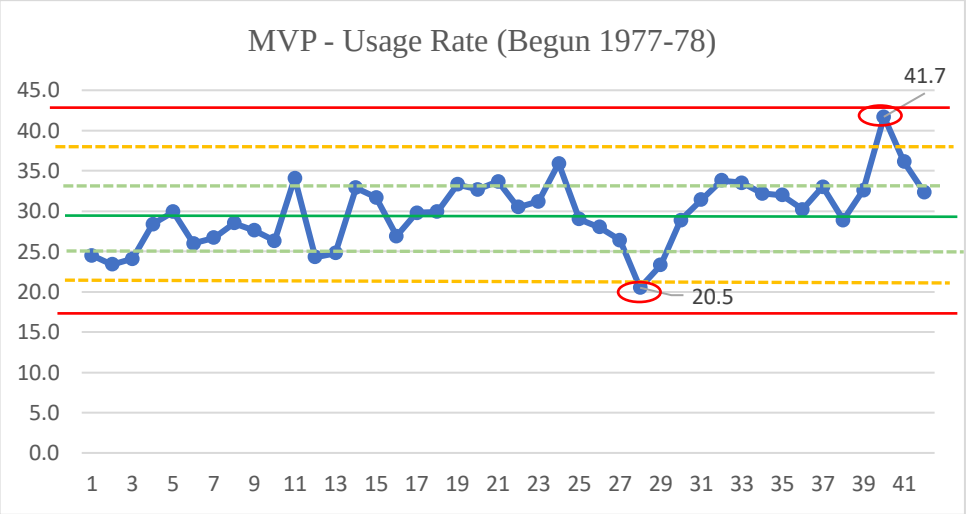

FIGURE 4. 10 USG Rate of MVP Recipients

He also had outstanding teammates in Amar'e Stoudmire, Shawn Marion, and Joe Johnson who carried more of the scoring load and other contributions such as rebounding and defending. Steve Nash played within a system and played his role; the team did not have to require a high usage rate from him to succeed. Conversely, there is Russell Westbrook who has the highest usage rate of all the MVPs throughout history. After Kevin Durant departed Oklahoma City, Westbrook took over the complete load on the team which led him to averaging a triple double throughout the season. Every play ran through him while he was on the court and he was a force on defense as well as in rebounding. Players who usually have high production tend to have a higher usage rate. As long as the player is efficient with their production, a higher usage rate tends to be beneficial.

Lastly, one of the most essential statistics in basketball is Player Efficiency Rating (PER). This variable measures a player's per-minute performance, while adjusting for pace.

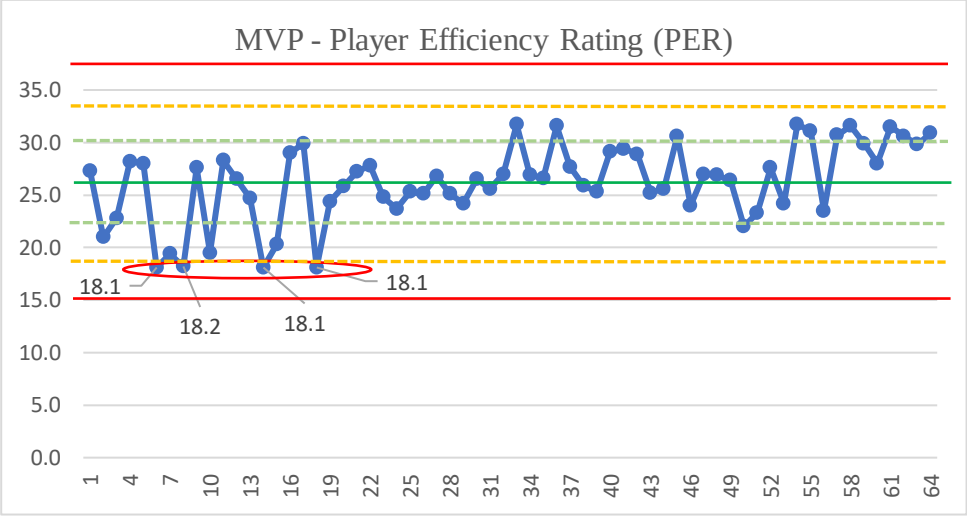

FIGURE 4. 11 PER of MVP Recipients

A league-average PER is approximately 15.00. The four outliers in FIGURE 4.11 include Bill Russell (1960-61, 1962-63) at 18.1 and 18.2, Wes Unseld (1968-69) at 18.1, and Dave Cowens (1972-73) at 18.1. Bill Russell was not a dominate offensive player as his TS% was below 50% each of the two years listed above. PER is highly swayed by efficiency on the offensive side of the ball, and Bill Russell strengths came as a defender and rebounder playing for a dominant team in the Boston Celtics. Wes Unseld was also not a big threat on the offensive end of the ball; not an exceptional scorer nor did he average many AST. He was most productive when on defense and in rebounding as well. His performance as a rookie was still impactful given the team's improvement from the year before. Likewise, Dave Cowens was not a strong player on offense like Bill Russell and Wes Unseld. His TS% was under 50% and his OWS were only 2.1 the year he won MVP. However, he led the league in DWS at 9.9 and also ranked top four in the league in MP and REB. This led the Boston Celtics to a 68-14 finish and the best record in the league. Player Efficiency Rating was not a sought out statistic in determining the most valuable player early in the league's history for quite a few players won the award by playing their role on the team. In the most recent years, MVP recipients have had a PER of around 25 or more, meaning they contribute more efficiently on the offensive end than players in the past.

4.3 Algorithm

The statistics that were introduced above will be used to form the proposed algorithm of this thesis. A multitude of variable combinations were used to create several formulas in hopes of determining an accurate prediction of the PTR of NBA players. Before formulating the algorithm

of this thesis, the methodology from the article, NBA MVP Prediction Model [20] was reproduced and used as a basis for comparison. The model from this article derived from the ideal that a player's value is based on their contribution to team wins and overall individual output. With that being said, the equation for player's value is provided below.

$$\text{Value} = 0.5(\text{Win Contribution}) + .\, 0.5(\text{Total Stats})$$

EQUATION 4. 1 Player Value

Win contribution is characterized as a player's level of impact and quality of impact. Level of impact measures how much a team relies on a player per possession compared to his teammates. The notable variables for level of impact are usage rate and wins, they depict the percentage of offensive plays used by a player and the influence they have on team wins or losses. Quality of impact measures the degree of excellence of a player's performance.

$$\text{Win Contribution} = \text{Level of Impact} * \text{Quality of Impact}$$

EQUATION 4. 2 Win Contribution

$$\text{Level of Impact} = (\text{Team Wins} * {}^{\text{Games Played}}/_{82} * {}^{\text{Minutes}}/_{48} * {}^{\text{Usage Rate}}/_{100}$$

EQUATION 4. 3 Level of Impact

$$\text{Quality of Impact} = 0.4(\text{VORP} + \text{Win Share}) + 0.2(\text{Net Rating})$$

EQUATION 4. 4 Quality of Impact

Lastly, the author simplified their equation for total stats. This consisted of statistical categories that the players accumulated through the entire season both on offense and defense. The author also scaled the equation so that it could be close in value to win contribution while providing optimum accuracy.

$$Total\ Stats = \frac{(Points * True\ Shooting\% + 1.5(Assists) + 1.2(Rebounds) + 3(Block) + 3(Steals) - Fouls - Turnovers}{25}$$

EQUATION 4. 5 Total Stats

When analyzing the results presented in the article, the model seemed to be fairly accurate. There were only two occasions out of eleven attempts where the actual MVP did not rank first; 2010-11 with LeBron James predicted to win over Derrick Rose and 2018-19 with James Harden winning over Giannis Antetokounmpo. Although the actual MVP did not rank first in both those seasons, the model did rank them second. It should also be mentioned that the predicted winners had compelling seasons which could make for a respectable argument of who is most deserving of the MVP honor. This study also revealed that players with high usage and low efficiency were undervalued in the model, big men forecasted values seemed to score higher, and defensive statistics were undervalued characteristics for the model. Using the same methodology, this

model was reproduced using statistics from players selected as an All-Star for the 2019-20 season. The results are shown in the table below. The table shows Giannis Antetokounmpo, the actual MVP of the 2019-20 season ranked second in player value behind James Harden. Also, the top four players in actual player value did fall within the top six of the predicted values with minor inaccuracies. Furthermore, Jimmy Butler had the greatest delta out of all the players, ranking 25th from the forecasted values when in actuality he had the 11th best player value during the 2019-20 season. Given these points, this model provides an adequate prediction, but embodies some inaccuracy.

TABLE 4. 7 Reproduced Model Forecast Results

Player	Value	Rank	Actual
James Harden	2988.10	1	3
Giannis Antetokounmpo	2894.01	2	1
LeBron James	2422.38	3	2
Anthony Davis	1980.02	4	6
Luka Doncic	1677.46	5	4
Nikola Jokic	1617.56	6	9
Damian Lillard	1492.13	7	8
Kawhi Leonard	1467.69	8	5
Jayson Tatum	1403.12	9	12
Khris Middleton	1342.54	10	-
Bam Adebayo	1097.67	11	-
Chris Paul	1079.46	12	7
Kyle Lowry	1011.15	13	-
Pascal Siakam	1001.71	14	10
Donovan Mitchell	1000.16	15	-
Devin Booker	934.70	16	-
Rudy Gobert	874.04	17	-
Kemba Walker	773.85	18	-
Russell Westbrook	665.37	19	-
Joel Embiid	533.44	20	-
Domantas Sabonis	517.04	21	-
Ben Simmons	397.04	22	-
Brandon Ingram	374.91	23	-
Trae Young	235.28	24	-
Jimmy Butler	16.27	25	11

The three techniques (ANN, KNN, LRM) were utilized to forecast the Regular Season MVP for the NBA using the programming platform, R Studio. Additionally, seven formulas were created using datasets of all the past MVP winners and runner up candidates. The seven formulas

were produced based on varying concepts in determining player value in basketball. The dependent variable used throughout all the formulas is PTR, which is the Total Voting Points (TVP) over the Voting Point Max (VPM). The formulas were used to create the models that correspond with it (e.g., formula one ~ model one) and then applied to the three techniques. The data used to generate the first six models were based off all the past MVP recipients. These models forecasted the PTR of the actual MVPs and the top five runner-ups from the 2009-2019 seasons. Studying the initial results of the first six models led to producing formula seven, which was intended to be the most accurate out of all the models. Consequently, model seven used data of the past MVPs and the runner-ups from the years 2009 to 2016 as it's foundation. Model seven was then used to forecast the PTR for the MVP and Runner-ups for the years 2016 to 2019. Before defining formula seven, the correlation between each of the variables and the PTR was calculated. The Kendall's tau coefficient was used to perform the correlation test. In the findings, the strongest variables were PER, VA, EWA, WS, WS48 and PM for they had the lowest p-values. By the same token, there were some other variables that were marginally strong as well that could also deem to be beneficial in executing the algorithm.

The first formula is the base statistics season total. The variables are all offensive statistics that a player accumulates throughout a season.

$$\mathrm{PTR} = \mathrm{PTS} + \mathrm{REB} + \mathrm{AST} + \mathrm{TwoPoint} + \mathrm{ThreePoint}$$

EQUATION 4. 6 Formula 1

The risk with this formula is that the most valuable player does not need to have the highest totals through the regular season to win the MVP award. This is not the most accurate formula because totals can tell a different story in comparison to a game by game stat-line. Totals can be manipulated by injuries, games played, minutes per game, etc. To better enhance this formula, defensive statistic totals should be included. Additionally, MPG and/or GP could be added to the formula so production by the amount of playing time gets captured. As currently constructed, the values may be assumed to be based on a full 82 game season, whereas in reality, a lot of players do not play a complete season. Minutes also impact the weight of the values, for example, two players can produce the same averages, but one may have better production because they can produce those numbers within a shorter timeframe.

The second formula represents statistics per game. This formula is composed of both offensive and defensive attributes, and it takes the averages of each variable from the entire regular season. This formula provides a depiction of the type of performance (statistical output) you'll get from a player per game. To further improve this formula, MPG and/or GP are variables that could be added with the same reasoning as formula one.

$$\mathrm{PTR} = \mathrm{PPG} + \mathrm{TRB} + \mathrm{APG} - \mathrm{TO} + \mathrm{SPG} + \mathrm{BPG}$$

EQUATION 4. 7 Formula 2

Formula three consist of advanced statistics. Advanced statistics refers to the exploration of complex multivariate relationships among variables. These statistics are used to determine

player and team value, determine trades, draft picks, team movement, etc. In this formula, some key statistics were used based on shooting and efficiency. To further expand this formula, MPG and/or GP could be included.

$$\mathrm{PTR} = \mathrm{FG} + \mathrm{TS} + \mathrm{VA} + \mathrm{EWA} + \mathrm{PER} + \mathrm{THREEFG} + \mathrm{FT}$$

EQUATION 4. 8 Formula 3

Formula four is a mixed formula of offensive averages and advanced statistics. The purpose of this formula is to determine how much of an impact offense has on MVP voting. To expand this formula, offensive rebounds and any other offensive attributes could be added, as well as MPG and/or GP.

$$\mathrm{PTR} = \mathrm{PPG} + \mathrm{APG} - \mathrm{TO} + \mathrm{FG} + \mathrm{ThreeFG} + \mathrm{FT} + \mathrm{USG} + \mathrm{OWS} + \mathrm{OPM}$$

EQUATION 4. 9 Formula 4

Formula five is a defensive mix formula with averages and advanced statistics. The aim of this formula is to determine how much of an impact defense has on MVP voting. To optimize this formula, defensive rebounds and any other defensive attributes could be added, as well as MPG and/or GP.

$$\mathrm{PTR} = \mathrm{SPG} + \mathrm{BPG} + \mathrm{DWS} + \mathrm{DPM}$$

EQUATION 4. 10 Formula 5

Formula six is a mix formula consisting of team wins, team losses and advanced statistics. The intent with this formula was to capture the players usage and direct impact on their team's success. One way to improve the accuracy of this formula would be to add in more team statistics – team success is a key component in determining the MVP.

$$\mathrm{PTR} = \mathrm{TW} - \mathrm{TL} + \mathrm{GP} + \mathrm{MPG} + \mathrm{USG} + \mathrm{VA} + \mathrm{EWA}$$

EQUATION 4. 11 Formula 6

The results of the first six formulas were used to highlight the key variables that defined an MVP. Ultimately, the best formulas were three, four, and six. Using the same attributes provided from the NBA MVP Prediction Model [20] article (Win Contributions, Individual Statistics, and Team Success), the variables were grouped into one of the attributes. Dependent upon the technique, the weight of the variables varied in order to make the model as efficient as possible. The base formula for Win Contributions is shown below.

$$\begin{aligned}\text{Win Contributions} &= \mathrm{GP} + \mathrm{MP} + C_1\mathrm{USG} + C_2\mathrm{EWA} + C_3\mathrm{WS48} + C_4\mathrm{VORP} + C_5\mathrm{OWS} + C_6\mathrm{DWS} \\ &+ C_7\mathrm{OPM} + C_8\mathrm{DPM} + C_9\mathrm{VA} + C_{10}\mathrm{PER}\end{aligned}$$

EQUATION 4. 12 Win Contributions

Individual Statistics consist of seasonal averages of basic stat categories that a player accumulates per game. The one advanced statistic included in this equation was TS, which is

used to measures a player's efficiency at shooting the ball. Dependent upon the technique, the weight of the variables varies to make the model as efficient as possible. The base formula for Individual Statistics is shown below.

$$\text{Individual Statistics} = \mathrm{GP} + \mathrm{MP} + \mathrm{C_1TS} + \mathrm{C_2PPG} + \mathrm{C_3APG} + \mathrm{C_4TRB} + \mathrm{C_5SPG} + \mathrm{C_6BPG} - \mathrm{C_7TO} - \mathrm{C_8PF}$$

EQUATION 4. 13 Individual Statistics

The last attribute is Team Success, which simply represents the player's team's level of achievement through the regular season. The equation for this attribute encompasses the teams record during the regular season and the team's net rating. The weight of the net rating changed between each of the techniques to make the models as efficient as possible. The base formula for Team Success is shown below.

$$Team\ Success = TW - TL + C_1 NRtg$$

EQUATION 4. 14 Team Success

Though the weight varies between techniques, the overall formula was consistent. Thus, formula 7 is presented.

$$PTR = Win\ Contribution + Individual\ Statistics + Team\ Success$$

EQUATION 4. 15 Formula 7

4.4 Forecast Techniques

To achieve optimum accuracy of the algorithm, the formulas were applied to the three different forecasting techniques (LRM, KNN, ANN) in order to formulate the models and determine the champion method. For LRM, the first six formulas were applied to the lm() function to create the models. Once all the models were generated, they were able to forecast the PTR for Giannis Antetokounmpo's 2018-19 season, when he won his first MVP honors. The reason for forecasting Antetokounmpo's PTR was to test the accuracy of each model by analyzing their MAPE values. Knowing that the models were generated from the data of only the past MVP winners, it was interesting to see how accurate the models would predict PTR values for runner-up candidates. The LRM models were then used to forecast the PTR for the MVPs and top five runner-ups from 2009-2019. The results of each model, filtered by year, were ranked to determine if the actual MVP placed 1st each time. Formula seven was then applied to the LRM technique and the weights were altered to achieve optimum accuracy. For LRM model seven, the WS and PM were split between offense and defense to improve efficiency. Below displays the weighted equations used to generate LRM model seven.

Win Contribution

$$= \mathrm{GP} + \mathrm{MP} + (200 * \mathrm{USG}) + (0.0001 * \mathrm{EWA}) + (1000 * \mathrm{WS48}) + (1000 * \mathrm{VORP}) + \big((100 * \mathrm{OWS}) + (1000 * \mathrm{DWS})\big) + \big((100 * \mathrm{OPM}) + (1000 * \mathrm{DPM})\big) + \mathrm{VA} + (1000 * \mathrm{PER})$$

EQUATION 4. 16 LRM Win Contribution

Individual Statistics

$$= (1000 * \mathrm{TS}) + (1000 * \mathrm{PPG}) + (100 * \mathrm{APG}) + (1000 * \mathrm{TRB}) + \mathrm{SPG} + \mathrm{BPG} - (0.0001 * \mathrm{TO}) - (0.0001 * \mathrm{PF}) + \mathrm{GP} + \mathrm{MP}$$

EQUATION 4. 17 LRM Individual Statistics

$$\text{Team Success} = \mathrm{TW} - \mathrm{TL} + (150 * \mathrm{NRtg})$$

EQUATION 4. 18 LRM Team Success

To implement the next technique, K-Nearest Neighbor, the same process as LRM is being adopted. The pre-processing transformation used range to scale the data so that it could fall within range-bounds (distance), a two-element numeric vector specifying a closed interval. All models except for model seven were built the same, below shows the weighted variables used for KNN model seven.

Win Contribution

$$= \mathrm{GP} + \mathrm{MP} + (50 * \mathrm{USG}) + (50 * \mathrm{EWA}) + (1000 * \mathrm{WS48}) + (1000 * \mathrm{VORP}) + (50 * \mathrm{WS}) + (1000 * \mathrm{PM}) + \mathrm{VA} + (5000 * \mathrm{PER})$$

EQUATION 4. 19 KNN Win Contribution

Individual Statistics

$$= (5000 * \mathrm{TS}) + (1000 * \mathrm{PPG}) + (1000 * \mathrm{APG}) + (100 * \mathrm{TRB}) + \mathrm{SPG} + \mathrm{BPG} - \mathrm{TO} - \mathrm{PF} + \mathrm{GP} + \mathrm{MP}$$

EQUATION 4. 20 KNN Individual Statistics

$$\text{Team Success} = \mathrm{TW} - \mathrm{TL} + (100 * \mathrm{NRtg})$$

EQUATION 4. 21 KNN Team Success

Once the formulas were applied to the models, it followed the same steps as the previous LRM models in order to generate the PTR forecast of the MVPs and runner-ups. As a reminder, model seven was the only model that used more recent data of MVP recipients and the top five runner-ups from the years 2009-2015.

Lastly, ANN models were implemented the same as the first two techniques. After converting all the variables to a time series function, they were then placed into a dataset. To reduce and even eliminate data redundancy, the dataset was normalized using the Max-Min

Normalization Function. After the data was normalized, the neuralnet() function was used. Before running the code, the seed had to be set to 100; thus, specifying the initial value of the random number seed so that the same result would appear each time the model was ran. The hidden neurons and threshold varied between the models. The models were then used to generate PTR values. Below displays the weighted equations for ANN model seven.

$$\text{Win Contribution} = \text{GP} + \text{MP} + (1.5 * \text{USG}) + (1.5 * \text{EWA}) + (1.5 * \text{WS48}) + \text{VORP} + (3 * \text{WS}) + (3 * \text{PM}) + (1.5 * \text{VA}) + (3 * \text{PER})$$

EQUATION 4. 22 ANN Win Contribution

$$\text{Individual Statistics} = (1.5 * \text{TS}) + \text{PPG} + (0.5 * \text{APG}) + (0.5 * \text{TRB}) + \text{SPG} + \text{BPG} - \text{TO} - \text{PF} + \text{GP} + \text{MP}$$

EQUATION 4. 23 ANN Individual Statistics

$$\text{Team Success} = \text{TW} - \text{TL} + \text{NRtg}$$

EQUATION 4. 24 ANN Team Success

The best models from each of the techniques were combined to forecast the PTR of NBA players, in hopes of predicting the MVP of the 2019-20 season. The three combination methods included Simple Average (SA), Ordinary Least Squares (OLS), and Constrained Least Squares

(CLS). To avoid forecasting the PTR value of all the players in the NBA, the sample size was filtered to only consider players that were selected as an All-Star for that season. Reason being the only player that has won an MVP award and was not selected as an All-Star was Karl Malone in 1998-99. Of course, he was no All-Star because there was no All-Star game that season due to the lockout. In the 57 years of selecting All-Star candidates, the MVP was either in the game, or voted in [16].The sample size for this study included only the 25 players that were selected as an All-Star for the 2019-20 season. The combination models were then used to analyze these specific players regular season performance, successfully generating their PTR. The optimum model from the set was then deemed to be the model used for the forecasting algorithm.

RESULTS

5.1 LRM

The LRM models one through six were used to predict the PTR of MVP recipients and the top five runner-ups for seasons 2009-2019. For LRM model seven, it forecasted the results of the MVP and top five runner-ups for the seasons 2016-2019. The analysis will show how the best model was selected when using LRM.

TABLE 5. 1 LRM Model Rankings

	1st place	2nd place	3rd place	Total
Model 1	60%	0%	20%	80%
Model 2	50%	10%	10%	70%
Model 3	50%	20%	20%	90%
Model 4	70%	20%	0%	90%
Model 5	50%	30%	20%	100%
Model 6	90%	10%	0%	100%
Model 7	100%	0%	0%	100%

Interestingly, in TABLE 5.1, all of the models had accurately ranked the forecast results of the actual MVP recipients first in PTR at 50% or greater. The strongest models proved to be models four, six, and seven which ranked the predicted values in first place at 70% or greater. These three models were the main focus of this continued analysis.

TABLE 5. 2 MAPE Value Range of Complete Dataset Using LRM

Measurement	#	Percentage	Sample Size
Very Accurate	24	6.35%	378
Accurate	32	8.47%	
Moderately Accurate	25	6.61%	
Slightly Accurate	27	7.14%	
Inexact	6	1.59%	
Out-of-Spec	264	69.84%	

The MAPE values indicated which range the forecast fell under dependent upon the value. The baseline data used as the foundation for the models caused the bulk of results to be out-of-spec. Models one through six were influenced by MVP caliber players only and had no cognizance of PTR values for players who had an average and/or subpar performance throughout the season. Disregarding the out-of-spec measurements, majority of the forecast were at least accurate.

TABLE 5. 3 MAPE Value Range of Actual MVP Recipients using LRM

Measurement	#	Percentage	Sample Size
Very Accurate	20	31.75%	63
Accurate	27	42.86%	
Moderately Accurate	12	19.05%	
Slightly Accurate	4	6.35%	
Inexact	0	0.00%	
Out-of-Spec	0	0.00%	

From TABLE 5.3, when filtering the forecast values to only account for the actual MVP recipients, it showed that the predicted results were at least accurate at approximately 75%. Notably, there were also no inexact or out-of-spec predictions either. This symbolized that using this technique will almost always provide a realistic prediction of the PTR for MVP candidates.

TABLE 5. 4 MAPE Value Range Filtered by LRM of Complete Dataset

Measurement	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Total	60	60	60	60	60	60	18
Very Accurate	1	3	5	5	4	4	2
Accurate	6	5	4	4	5	5	3
Moderately Accurate	4	3	2	2	6	6	2
Slightly Accurate	4	5	5	6	1	3	3
Inexact	1	2	1	0	2	0	0
Out-of-Spec	44	42	43	43	*42*	42	8
Accuracy Percentage	11.67%	13.33%	15.00%	15.00%	15.00%	15.00%	27.78%

Favoring the three potential models, model seven had a significant boost in accuracy as seen in TABLE 5.4. Even though the sample size for model seven is smaller, it was at least a 12% gap in MAPE accuracy compared to the other models. Model four seemed to be the second best model.

TABLE 5. 5 MAPE Value Range Filtered by LRM of Actual MVP Recipients

Measurement	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Total	10	10	10	10	10	10	3
Very Accurate	1	3	5	5	2	3	1
Accurate	6	5	4	4	4	4	0
Moderately Accurate	2	2	0	1	4	3	0
Slightly Accurate	1	0	1	0	0	0	2
Inexact	0	0	0	0	0	0	0
Out-of-Spec	0	0	0	0	*0*	0	0
Accuracy Percentage	70.00%	80.00%	90.00%	90.00%	60.00%	70.00%	33.33%

Referring to TABLE 5.5, the low value in accuracy percentage for model seven was initially concerning, but then again, the sample size used for that model should not be ignored. After continuing to analyze the chart, model four had an accuracy of 90% and none of the predicted results were categorized as inexact or out-of-spec. Given the points made in this section, both LRM models four and seven could be used to generate accurate forecast for PTR.

5.2 KNN

Below presents the analysis conducted on all the KNN models. Models one through seven underwent a similar analysis to that of the LRM models.

TABLE 5. 6 KNN Model Rankings

	1st place	2nd place	3rd place	Total
Model 1	50%	20%	10%	80%
Model 2	30%	20%	10%	60%
Model 3	70%	10%	10%	90%
Model 4	70%	20%	0%	90%
Model 5	40%	30%	10%	80%
Model 6	70%	10%	10%	90%
Model 7	67%	33%	0%	100%

TABLE 5.6 shows the ranking results of the forecast values for each model filtered by year when using KNN. The strongest models from the table were models three, four and six, ranking the PTR of the MVPs in first place at 70%. Not to mention, all three of the models ranked the MVP recipients in the top three at 90%. These three models were the main focus of the continued analysis.

TABLE 5. 7 MAPE Value Range of Complete Dataset Using KNN

Measurement	#	Percentage	Sample Size
Very Accurate	17	4.50%	378
Accurate	36	9.52%	
Moderately Accurate	30	7.94%	
Slightly Accurate	23	6.08%	
Inexact	6	1.59%	
Out-of-Spec	266	70.37%	

Analyzing the accuracy from TABLE 5.7, there is an increase of out-of-spec measurements in comparison to the LRM models. Disregarding the out-of-spec measurements, the percentage of at least accurate forecast results were about 14%. Just from interpreting this table, an assumption was made that this technique was decent but not as strong as LRM. By all means, this was not yet proven and required further analysis.

TABLE 5. 8 MAPE Value Range of Actual MVP Recipients using KNN

Measurement	#	Percentage	Sample Size
Very Accurate	14	22.22%	63
Accurate	34	53.97%	
Moderately Accurate	10	15.87%	
Slightly Accurate	1	1.59%	
Inexact	2	3.17%	
Out-of-Spec	2	3.17%	

The most compelling evidence in suggesting the strength of KNN is shown in TABLE 5.8, where majority of the predicted values of the actual MVP recipients were accurate. The table above shows that this technique will provide an accurate forecast for the most valuable player just over 76%. However, LRM again proved to be better because it did not have any forecast values in the inexact nor out-of-spec range for MVP recipients.

TABLE 5. 9 MAPE Value Range Filtered by KNN of Complete Dataset

Measurement	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Total	60	60	60	60	60	60	18
Very Accurate	4	2	2	2	3	4	0
Accurate	5	7	7	6	3	6	2
Moderately Accurate	1	6	4	3	8	6	2
Slightly Accurate	7	2	3	5	2	2	2
Inexact	0	1	2	1	1	0	1
Out-of-Spec	43	42	42	43	*43*	42	11
Accuracy Percentage	15.00%	15.00%	15.00%	13.33%	10.00%	16.67%	11.11%

Focusing on KNN models three, four and six; it can be seen that model six and model three provided more accurate predictions after observing TABLE 5.9. Consequently, KNN model four was removed from consideration.

TABLE 5. 10 MAPE Value Range Filtered by KNN of Actual MVP Recipients

Measurement	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Total	10	10	10	10	9	10	3
Very Accurate	4	1	2	2	2	3	0
Accurate	5	7	7	6	3	5	1
Moderately Accurate	1	2	1	1	3	2	0
Slightly Accurate	0	0	0	0	0	0	1
Inexact	0	0	0	1	1	0	0
Out-of-Spec	0	0	0	0	*0*	0	1
Accuracy Percentage	90.00%	80.00%	90.00%	80.00%	55.56%	80.00%	33.33%

Referencing TABLE 5.10, model three proved to be slightly more accurate than model six. Since model three deemed to be consistent throughout this analysis, it can be suggested that it is the best model to be used for KNN.

5.3 ANN

Similar to the previous methods, the formulas corresponded with the models and are analyzed the same. Only difference is that a combination model using ANN models one through six was included. This section highlights the results of the ANN models and depicts the most efficient model to use under this technique.

TABLE 5. 11 ANN Model Ranking

	1st place	2nd place	3rd place	Total
Model 1	30%	20%	10%	60%
Model 2	0%	30%	30%	60%
Model 3	50%	20%	10%	80%
Model 4	20%	40%	0%	60%
Model 5	20%	0%	20%	40%
Model 6	70%	10%	20%	100%
Combo (1-6)	40%	20%	20%	80%
Model 7	66.67%	0.00%	33.33%	100%

The ranking results of each model filtered by year of the actual MVP recipients are shown in TABLE 5.11. Certainly, the greater the percentage of first place rankings signifies better accuracy for that model. If the PTR of the actual MVP recipient ranks within the top three, it may be a compelling model that could provide an accurate prediction. Referencing the table above, Model six was the strongest out of the bunch where 70% of the actual MVP recipients ranked first in PTR. The other 30% of MVP recipients for Model six was distributed between second and third place. This model provided the most accurate results using ANN in regard to player ranking.

TABLE 5. 12 MAPE Value Range of Complete Dataset Using ANN

Measurement	#	Percentage	Sample Size
Very Accurate	24	5.48%	438
Accurate	28	6.39%	
Moderately Accurate	30	6.85%	
Slightly Accurate	34	7.76%	
Inexact	17	3.88%	
Out-of-Spec	305	69.63%	

TABLE 5.12 gives a depiction of how accurate each forecast was when using ANN. Disregarding the out-of-spec measurements, majority of the forecasted values were slightly and moderately accurate. With this in mind, ANN may be one of the weaker forecasting methods.

TABLE 5. 13 MAPE Value Range of Actual MVP Recipients using ANN

Measurement	#	Percentage	Sample Size
Very Accurate	20	27.40%	73
Accurate	22	30.14%	
Moderately Accurate	13	17.81%	
Slightly Accurate	8	10.96%	
Inexact	2	2.74%	
Out-of-Spec	8	10.96%	

TABLE 5.13 reflected that ANN could provide accurate predictions of the actual MVPs. On the other hand, it revealed that the forecast for runner-up candidates were inaccurate given the vast differences in results when comparing to TABLE 5.12. Variables such as luck and injury were not considered, but TABLE 5.13 proved that the PTR values of the actual MVPs were reliable predictions.

TABLE 5. 14 MAPE Value Range Filtered by ANN of Complete Dataset

Measurement	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Combo Model	Model 7
Total	60	60	60	60	60	60	60	18
Very Accurate	1	4	2	5	2	6	1	3
Accurate	3	4	7	3	2	4	4	1
Moderately Accurate	5	2	3	3	4	4	9	0
Slightly Accurate	6	7	2	3	6	3	4	3
Inexact	2	4	5	2	1	1	1	1
Out-of-Spec	43	39	41	44	45	42	41	10
Accuracy Percentage	6.67 %	13.33 %	15.00 %	13.33 %	6.67 %	16.67 %	8.33 %	22.22 %

In TABLE 5.14, ANN model seven provided the best accuracy percentage, but the sample size is much smaller. Instead, the most promising model was ANN model six with an accuracy percentage of 16.67%. Keep in mind, ANN model six was the best ranking model used under this technique.

TABLE 5. 15 MAPE Value Range Filtered by ANN of Actual MVP Recipients

Measurement	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Combo Model	Model 7
Total	10	10	10	10	10	10	10	3
Very Accurate	1	3	2	4	1	5	1	2
Accurate	3	3	6	2	2	4	3	0
Moderately Accurate	4	0	2	2	1	0	4	0
Slightly Accurate	2	3	0	2	0	0	1	0
Inexact	0	0	0	0	1	0	1	0
Out-of-Spec	0	1	0	0	*5*	1	0	1
Accuracy Percentage	40.00 %	60.00 %	80.00 %	60.00 %	30.00 %	90.00 %	40.00 %	66.67 %

TABLE 5.15 analyzes the MAPE values of the actual MVP recipients filtered between each of the ANN models. As suspected, model six provided the highest accuracy percentage, thus validating that it was the most accurate model under ANN.

5.4 MVP Algorithm

TABLE 5. 16 Favored Models Comparison

Model	MAPE Point System	Actual MVP Forecasted Ranking
LRM Model 4	15	2
LRM Model 7	20	1
KNN Model 3	10	15
ANN Model 6	13	11

Utilizing the best models from each of the techniques (LRM Model 4, LRM Model 7, KNN Model 3 and ANN Model 6), predictions were forecasted of the PTR for the Most Valuable Player for the 2019-20 season. Following TABLE 5.16, LRM models four and seven proved to have the best results. Additionally, the rankings were also promising, placing Giannis second using LRM Model four and first when using LRM Model seven. LRM proved to be the best technique out of the three that were presented in this thesis given the data results. Since both LRM models had promising results, these models were combined to create a more efficient model.

TABLE 5. 17 Combination Models Comparison

	MAPE Point System	Actual MVP Forecasted Ranking
SA	15	1
OLS	5	1
CLS	10	1

There were three combination methods (SA, OLS, CLS) used in finalizing the MVP forecast algorithm. These combination models were compared for accuracy and repeatability. Interestingly enough, all three models ranked Giannis Antetokounmpo (2019-20 MVP) first in PTR of the 2019-20 season. Regarding the forecasted values, OLS was most inaccurate, so this model was dismissed as a prediction method. For SA and CLS, the forecasted values resembled accurate results. The actual top six players were ranked one through six (not in order) in the forecasted results. This verified that these two combination models would provide an accurate prediction for MVP; and predict the top three to six players in PTR of a regular season. Since SA

had a better MAPE accuracy, this would be the base model for the algorithm; however, the CLS model could be used in the algorithm as a verification method.

TABLE 5. 18 Forecast Algorithm Results

Player	Actual Ranking	SA Forc Combo	SA Ranking	CLS Forc Combo	CLS Ranking	OLS Forc Combo	OLS Ranking
Giannis Antetokounmpo	MVP	0.969	1	0.990	1	0.637	1
LeBron James	1st Runner-up	0.672	4	0.488	5	0.334	4
James Harden	2nd Runner-up	0.769	2	0.604	2	0.423	2
Luka Doncic	3rd Runner-up	0.693	3	0.504	4	0.351	3
Kawhi Leonard	4th Runner-up	0.645	6	0.474	6	0.312	6
Anthony Davis	5th Runner-up	0.653	5	0.513	3	0.325	5
Chris Paul	6th Runner-up	0.270	21	-0.090	20	-0.057	22
Damian Lillard	7th Runner-up	0.508	7	0.141	11	0.153	7
Nikola Jokic	8th Runner-up	0.444	11	0.170	10	0.114	11
Pascal Siakam	9th Runner-up	0.377	13	0.066	13	0.047	13
Jimmy Butler	10th Runner-up	0.325	15	-0.020	17	-0.005	15
Jayson Tatum	11th Runner-up	0.470	8	0.182	9	0.134	8
Devin Booker	-	0.290	18	-0.106	22	-0.046	21
Domantas Sabonis	-	0.285	19	-0.032	19	-0.035	17
Ben Simmons	-	0.239	24	-0.125	23	-0.085	23
Russell Westbrook	-	0.402	12	0.118	12	0.075	12
Bam Adebayo	-	0.279	20	-0.014	15	-0.035	18
Joel Embiid	-	0.460	9	0.183	8	0.127	10
Rudy Gobert	-	0.261	22	0.014	14	-0.042	20
Brandon Ingram	-	0.223	25	-0.236	24	-0.118	25
Kyle Lowry	-	0.317	16	-0.022	18	-0.011	16
Khris Middleton	-	0.455	10	0.219	7	0.132	9
Donovan Mitchell	-	0.295	17	-0.098	21	-0.041	19
Kemba Walker	-	0.354	14	-0.017	16	0.015	14
Trae Young	-	0.252	23	-0.269	25	-0.104	24

5.5 Discussion of Results

In this study, an algorithm was formed to accurately determine the Most Valuable Player of each season. The optimum strategy that efficiently uses player/team statistics to predict the PTR metric is the Simple Average Linear Regression Model (SA-LRM) using formulas four and seven. The Constrained Least Squares Linear Regression Model (CLS-LRM) will also be utilized in the algorithm to validate the results from the SA-LRM. As stated before, one cannot be considered an MVP candidate if they were not selected as an All-Star. Some players may have moments in the season where their performance is highly productive and entertaining. However, to be a regular season MVP, a player must be consistently great throughout the entirety of the season. There are many factors that go into determining the MVP, the three factors analyzed in this study were individual statistics, win contributions, and team success. These were thought to be impactful variables that alludes to a player's value, which can be represented as quantitative data. In opposition, qualitative data is more difficult to analyze, factors such as physical condition, psychological state, individual characteristics, etc. were not included in this study; but could have an effect on the decision of MVP. Using the preferred algorithm, only quantitative data will be needed to make an accurate prediction.

When formulating the algorithm, there were many out of spec forecast values, meaning the predicted values were not close at all to the actual value. The first reason for the out-of-spec measurements, is due to the data used to create the models. The statistics were based on high performance players, thus influencing the forecasted results. If the data used for the foundation

of the models was a mix of high to low performance players, the overall forecast values may be more accurate. Adopting the idea of using All-Star players to be used for the sample size in order to evaluate the models deemed beneficial, for these are the players that are usually voted on for MVP. Some of the All-Stars do not receive any votes for MVP, which means their data could be included in forming the foundation for the forecasting algorithm in order to provide more accurate PTR results for all players. Analyst should caution using outdated statistics as the foundation of their models as well. The game of basketball is continuously changing and evolving, and the way the game was played back in the 60s is completely different to how it is played today. Using data from more recent players would be suggested and give a more accurate depiction.

Through this study, an algorithm was formed that could accurately depict the MVP of a regular season. This algorithm can be utilized once the All-Star teams are set, and the players have been selected. It can be used to track the players who are most deserving of the award without having to conduct a deep analysis into players statistics. Individuals who cast a vote for MVP could use this algorithm as guidance to who they should vote for. It could also be used as entertainment for the fans, to discuss the rankings and debate about who will win. This study provided great insight about what it takes to become an MVP. For most, it was no easy task, having to carry a team and lead them in multiple statistics in an effort to bring success to the franchise. Whereas some players are put into the right system that will showcase their talents and provide them with vast opportunities. This algorithm will be able to predict the percentage of

votes a player deserves given their performance. The player with the highest PTR percentage will be considered the most valuable player.

CONCLUSIONS

The purpose of this study was to create an algorithm that could be used during every season that could accurately predict the MVP. There were seven underlying models used between three different techniques to determine the optimum model that could be utilized to forecast the Most Valuable Player. After the initial analysis, Linear Regression Model was found to be the most promising technique among the three being studied. Four underlying models were found to be effective dependent upon the technique being used. The best four pairs of underlying models and techniques were then used to forecast the PTR of the players selected to be an All-Star for the 2019-20 season in hopes of determining the best model out of the four. Both of the LRM models provided accurate results which led to combining the forecasts to improve the accuracy. The combination methods used were Simple Average, Ordinary Least Squares, and Constrained Least Squares to find the best combination forecast. The results showed that all three methods ranked the players correctly, but in terms of accurate value, Simple Average was the most accurate. It must be noted that Constrained Least Squares could also be used to verify the findings from Simple Average. The Simple Average combination of the forecasts from two linear regression models can be used each year from the final selection of All-Stars until the end of the season. The algorithm also tracks the players who are most deserving of the MVP accolade.

To improve this algorithm for future works, the first suggestion would be to use percentage variables instead of totals and/or averages. Percentage variables captures the

likelihood of a player succeeding in that particular field. This study would have been even more accurate had the percentage variables been utilized. Another area of improvement would be to utilize different methods of data collection, such as motion capture technologies which allows one to track every team or player movement on the court in milliseconds to obtain physical and/or emotionally variables that could have an effect on MVP voting. If these alternative methods were used, then more qualitative data could be captured. While three forecasting techniques have been implemented in this study, investigating other techniques may further improve the algorithm. Determining the right players to pull data from and build the model foundation is vital. Instead of using data points from past MVP players and the runner-ups, past All-Star selectees could be a bigger pool for future studies. Additionally, the sample size should increase so that the algorithm can be deemed repeatable. Overall, the algorithm developed from this study produces accurate forecasts of MVP and could be innovated to improve accuracy and/or predict other aspects in the great game of basketball.

REFERENCES

[1] L. Steinberg, "CHANGING THE GAME: The Rise of Sports Analytics," *Forbes*, 18-Aug-2015. [Online]. Available: https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/?sh=1dc21b3b4c1f. [Accessed: 04-Jan-2021].

[2] J. J. Cochran, "The Emergence of Sports Analytics," *Sports Analytics*, Feb. 2010.

[3] F. Media, Ed., "Dr. James Naismith's Life," *Naismith Basketball Foundation*, 13-Nov-2014. [Online]. Available: https://naismithbasketballfoundation.com/james-naismith-life/. [Accessed: 04-Jan-2021].

[4] C. Martinez, "The first intercollegiate basketball game was played on Feb. 9, 1895," *NCAA.com*, 09-Feb-2017. [Online]. Available: https://www.ncaa.com/news/basketball-men/article/2016-02-09/possible-first-intercollegiate-basketball-game-was-played-feb. [Accessed: 04-Jan-2021].

[5] "NBA History.," *NBAHOOPSONLINE.com: NBA History.*, 2015. [Online]. Available: https://www.nbahoopsonline.com/. [Accessed: 04-Jan-2021].

[6] C. Gough, "Share of players in the NBA from 2010 to 2020, by ethnicity." Jul-2020.

[7] M. Corvo, "How Voting is Done for the NBA MVP and Its Evolution," *ClutchPoints*, 21-Feb-2020. [Online]. Available: https://clutchpoints.com/how-voting-is-done-for-the-nba-mvp-and-its-evolution/. [Accessed: 04-Jan-2021].

[8] M. Olsofka, "Why is Data Analytics So Important in Sports?," *Samford University*, 08-Aug-2018. [Online]. Available: https://www.samford.edu/sports-analytics/fans/2018/Why-is-Data-Analytics-So-Important-in-Sports. [Accessed: 06-Jan-2021].

[9] V. Sarlis and C. Tjortjis, "Sports analytics — Evaluation of basketball players and team performance," *Information Systems*, vol. 93, p. 101562, 2020.

[10] Y. Chen, J. Dai, and C. Zhang, "A Neural Network Model of the NBA Most Valued Player Selection Prediction," *Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence - PRAI '19*, 2019.

[11] C. Clabaugh, D. Myszewski, and J. Pang, Eds., "Neural Network Website," *Neural Networks - History*. [Online]. Available: https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/history1.html. [Accessed: 06-Jan-2021].

[12] M. Khashei and M. Bijari, "An artificial neural network (p,d,q) model for timeseries forecasting," *Expert Systems with Applications*, vol. 37, no. 1, pp. 479–489, Jan. 2010.

[13] L. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[14] S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," *S B Imandoust et al. Int. Journal of Engineering Research and Applications*, vol. 3, no. 5, 2013.

[15] D. Kopf, "The Discovery of Statistical Regression," *Priceonomics*, 06-Nov-2015. [Online]. Available: https://priceonomics.com/the-discovery-of-statistical-regression/. [Accessed: 07-Jan-2021].

[16] A. Alvarez, "Which NBA players, if any, won the regular season MVP award but were not selected for the All-Star game?," *Quora*, 19-Jul-2013. [Online]. Available: https://www.quora.com/Which-NBA-players-if-any-won-the-regular-season-MVP-award-but-were-not-selected-for-the-All-Star-game. [Accessed: 07-Jan-2021].

[17] G. Vinué and I. Epifanio, "Forecasting basketball players' performance using sparse functional data*," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 12, no. 6, pp. 534–547, Sep. 2019.

[18] A. W. Nutting, "Individual Tournament Incentives in a Team Setting: The 2008-09 NBA MVP Race," International Journal of Sport Finance; vol. 5, no. 3, pp. 208–221, Aug. 2010.

[19] M. Yang, Y. Wei, L. Liang, J. Ding, and X. Wang, "Performance evaluation of NBA teams: A non-homogeneous DEA approach," Journal of the Operational Research Society, pp. 1–12, Feb. 2020.

[20] P. Li, "NBA MVP Prediction Model," Medium, 20-Apr-2019. [Online]. Available: https://towardsdatascience.com/nba-mvp-predictor-c700e50e0917. [Accessed: 08-Jan-2021].

[21] C. Brighenti, "Using Data Science to Predict the Next NBA MVP," Medium, 22-Jul-2019. [Online]. Available: https://towardsdatascience.com/using-data-science-to-predict-the-next-nba-mvp-30526e0443da. [Accessed: 08-Jan-2021].

[22] S. Wu, “NBA MVP Prediction Model,” Medium, 03-Nov-2019. [Online]. Available: https://medium.com/@suriwu2019/nba-mvp-prediction-model-af4a55bcd8b7. [Accessed: 08-Jan-2021].

[23] W. Tichy, “Changing the Game: Dr. Dave Schrader on sports analytics,” Ubiquity, vol. 2016, no. May, pp. 1–10, May 2016.

[24] B. J. Coleman, “Identifying the ‘Players’ in Sports Analytics Research,” Interfaces, vol. 42, no. 2, pp. 109–118, 2012.

[25] L. Sha, P. Lucey, Y. Yue, X. Wei, J. Hobbs, C. Rohlf, and S. Sridharan, “Interactive Sports Analytics: An Intelligent Interfacefor Utilizing Trajectories for Interactive Sports PlayRetrieval and Analytics,” ACM Transactions on Computer-Human Interaction, vol. 25, no. 2, pp. 1–32, Apr. 2018.

[26] E. Morgulev, O. H. Azar, and R. Lidor, “Sports analytics and the big-data era,” International Journal of Data Science and Analytics, vol. 5, no. 4, pp. 213–222, Aug. 2018.

[27] R. M. Musa, A. P. P. A. Majeed, Z. Taha, M. R. Abdullah, A. B. H. M. Maliki, and N. A. Kosni, “The application of Artificial Neural Network and k-Nearest Neighbour classification models in the scouting of high-performance archers from a selected fitness and motor skill performance parameters,” Science & Sports, vol. 34, no. 4, Sep. 2019.

[28] L. Zylstra, “NBA: The Four Different Definitions of MVP, and the obvious solution,” Medium, 22-Oct-2018. [Online]. Available: https://lzylstra.medium.com/nba-the-four-different-definitions-of-mvp-and-the-obvious-solution-68a86eb15ca4. [Accessed: 05-Feb-2021].

[29] J. Brownlee, “Linear Regression for Machine Learning,” Machine Learning Mastery, 25-Mar-2016. [Online]. Available: https://machinelearningmastery.com/linear-regression-for-machine-learning/. [Accessed: 10-Feb-2021].

[30] A. Ricky, “How Data Analysis In Sports Is Changing The Game,” Forbes, 31-Jan-2019. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2019/01/31/how-data-analysis-in-sports-is-changing-the-game/?sh=3147492b3f7b. [Accessed: 21-Feb-2021].

[31] D. Kopf, “Data analytics have made the NBA unrecognizable,” Quartz, 18-Oct-2017. [Online]. Available: https://qz.com/1104922/data-analytics-have-revolutionized-the-nba/. [Accessed: 21-Feb-2021].

APPENDIX

APPENDIX A: Datasets

- Statistical Dataset of MVP Recipients in the NBA:
  https://www.kaggle.com/jordanmccorey/statistical-dataset-of-mvp-recipients-in-the-nba
- Statistical Dataset of MVP Recipients and Runner-ups in the NBA:
  https://www.kaggle.com/jordanmccorey/statistical-dataset-of-mvpsrunnerups-in-nba
- Forecast Dataset of All Star Selectees:
  https://www.kaggle.com/jordanmccorey/forecast-dataset-of-all-star-selectees
- LRM Model Results:
  https://www.kaggle.com/jordanmccorey/lrm-models-16-results
  https://www.kaggle.com/jordanmccorey/lrm-model-7-results
- KNN Model Results:
  https://www.kaggle.com/jordanmccorey/knn-model-16-results
  https://www.kaggle.com/jordanmccorey/knn-model-7-results
- ANN Model Results:
  https://www.kaggle.com/jordanmccorey/ann-models-16-reults
  https://www.kaggle.com/jordanmccorey/ann-model-7

APPENDIX B: R Program Models

- LRM Forecasting Model:
  https://www.kaggle.com/jo rdanmccorey/lrm-forecasting-model-for-mvp-of-nba
- KNN Forecasting Model:
  https://www.kaggle.com/jordanmccorey/knn-forecasting-model-mvp-of-nba
- ANN Forecasting Model:
  https://www.kaggle.com/jordanmccorey/ann-forecasting-model-mvp-of-nba

- Reproduce Program
  https://www.kaggle.com/jordanmccorey/reproduce-program-mvp-of-nba

- Forecast 2019-20 MVP of NBA
  https://www.kaggle.com/jordanmccorey/forecast-2019-20-mvp-of-nba