**Springboard Capstone Project I Milestone Report**

**Predicting the Next NBA MVP**

**By Dillon Quan**

**The Problem**

Every year in sports, specifically basketball, many accolades will be given to players who are acknowledged for their individual achievements. One of the most honorable awards in the NBA is the Most Valuable Player (MVP) award. This award is usually given to an individual who has demonstrated consistent excellence and performance throughout the entire season. But, one might ask, how are players selected each year for this award? Well, there is an entire process. Before the 1980-1981 NBA season, the winner for the award was selected by NBA players. Now, the MVP is determined by a group of sports analysts and broadcasters where they each rate their top five candidates. Each candidate will get points depending on the rankings from the vote and the candidate with the most points will win the award. Although these analysts and broadcasters are professionals who know the sport inside out, one can argue this process may be biased or subjective. As a result, for a more objective way, the goal of my first capstone project will be to develop a model to predict future MVPs using a dataset that categorizes between previous MVPs and non MVPs.

**The Client**

The target audience this project aims to draw are sports analysts and anchors, sports analytic companies, the NBA league, NBA teams, and most of all, the fans. By developing a successful model to predict these prospects, anyone who has an interest in the sport may gain further insight on the discern trends from previous winners rather than the standard eye test. More importantly, the same classification methods can be used for other fields to classify other categories if desired. For instance, it can help credit collection companies classify their successful outbound calls or classify employees’ satisfaction with their company based on taken surveys. In the end, even though this project specifically gears towards sport enthusiasts, this process can be applicable to anyone who has a question that can be solved by classification.

**Data Wrangling**

To approach this problem, a dataset was retrieved from Kaggle that contains players’ stat lines from all seasons scrapped from basketball reference. After inspecting and understanding the dataset, I began cleaning and filtering it in my attempt to address this question. I started by dropping extra features that were present due to web scrapping. Then, the dataset was filtered to only contain players from more recent seasons, primarily 2000 – 2017. Since this is a class imbalanced problem, meaning there is a disparity between the number positive and negative labels, MVPs and non MVPs, each season will be limited to one hundred players for sampling which will consist of ninety-nine non MVPs and one MVP. Lastly, players that were traded during a season were also filtered out of the data for better sampling.

After the data was filtered, it was then cleaned for readability. For instance, each player stat line was converted from total season stats to per game stats. Then, a few additional features were added despite already containing many regular and advance stats. First, the target column feature that distinguishes between MVPs from non MVPs were included. This is crucial since the model will learn from these labels. Then, the Win – Loss Ratio and the marginal PER were also implemented onto the data. The Win – Loss Ratio determines the success of the team and the marginal PER calculates how well a player performed compared to the league average for that season. More of this will be discussed in the statistic inference section.

The Win – Loss Ratio was gathered from web scrapping basketball-reference.com. Since there were eighteen seasons of player data, this process was done eighteen times each time using a different link. The data retrieved from web scrapping was then exported into a csv file so it can merge with the main dataset.

**Data Story**

Despite cleaning up the dataset, it still isn’t known how these panel of pundits cast their votes. What traits do they look at when picking and ranking their candidates? Through research, majority of the consensus seems to gravitate towards consistent excellence and value an individual brings to his team. But how is this measured? Luckily, there are a few stats that are recorded for each player that help quantify what these things are. These are the Player Efficiency Rating (PER) and Win Shares (WS). The figures referenced in this section can be found in NBA Player EDA notebook page.

PER

To provide some insight on how this stat quantifies consistent excellence, further discussion on this metric is needed. PER is a single number that measures an individual’s positive and negative contribution on the court such as points, assists, rebounds, turnovers, fouls, etc. It is a metric that is adjusted for the minutes an individual played. This allows comparison between players who has a disparity in minutes. Also, this stat is pace adjusted, meaning players that play on a fast pace team won’t get a higher PER than players who plays on a slower pace team.

Strengths

-This metric consolidates almost all box stats recorded into one number. The league average PER is set to 15. This provides a benchmark to help rate each player’s accomplishments on the court and allows comparison between players.

Weaknesses

-Although this metric takes into account many box statistics, it fails to accurately quantify players who are defensive minded players that excel at playing individual or team defense. This is the case because PER only counts steals, blocks, and defensive rebound percentage for defense. Things like contesting shots or playing solid defense isn’t taking into account which, as a result, makes this stat more skewed towards offense.

However, in this case, since I am using PER to help determine MVPs and not Defensive Player of the Year (DPOY), this metric will be fine. It is certainly a plus if the MVP contributes immensely on the defensive end of the game, but the trends form previous winners of the award have shown that majority of the players that won the MVP award in the past two decades did not need to contribute greatly in this part of the game. This is shown through Defensive Win Shares. This will be discussed down below.

Basketball Reference (https://www.basketball-reference.com/about/per.html) provides a detailed explanation and the formula used to calculate this metric. The benchmark for this metric can be seen below.

All-time great season 35.0+

Runaway MVP candidate 30.0-35.0

Strong MVP candidate 27.5-30.0

Weak MVP candidate 25.0-27.5

Definite All-Star 22.5-25.0

Borderline All-Star 20.0-22.5

Second offensive option 18.0-20.0

Third offensive option 16.5-18.0

Slightly above-average player 15.0-16.5

Rotation player 13.0-15.0

Non-rotation player 11.0-13.0

Fringe roster player 9.0-11.0

Player who won't stick in the league 0-9.0

Win Share (WS)

This metric was developed to quantify a player’s individual contribution to team wins. Win share is a player statistic that attempts to divvy up credit for the team’s success to the individuals on the team. Some give three Win Shares per win while others give one Win Share per win. Since the dataset is obtained from Basketball Reference, the Win Shares column are calculated based on one win is equivalent to one Win Share. Win Share is also equal to the sum of both Offensive Win Share and Defensive Win Share. Basketball Reference (https://www.basketball-reference.com/about/ws.html) provides a more detailed explanation on how the metric is calculated.

Strengths

-This metric helps quantify each individual’s contributions that attributes to team success.

Weaknesses

-Since defensive possession requires a team effort, the Defensive Win Shares for players on the same team are relatively close to each other. This allow players who aren’t so good on defense have a slight increase in rating if they happen to play with great defensive players.

Since these two metrics are used to help quantify what it takes to be an MVP, a scatter plot is plotted with PER vs WS with only MVPs stat lines for the first figure. From the graph, it can be seen that there is a positive correlation between these two metrics. The figure shows that MVP caliber players attribute to approximately 15 wins alone on average.

In the second figure, Defensive Win Share vs Offensive Win Share is plotted using MVP stat lines. From this graph, it can be seen that the Offensive Win Share on average is about two to three times higher than Defensive Win Share. As a result, defense isn’t emphasized when choosing MVPs.

While exploring the MVPs stat lines, the following trends were identified. Most players win the MVP award between 24 – 30 years old. To the general viewer of the game, this age range falls under the range where players are approaching or already in their primes. It can also be seen from the categorical time series plot, the first decade had primarily big men (Center, Power Forward) winning the award. In the following decade, the award was given to primarily smaller, faster, and versatile players (Point Guard, Small Forward). There is one caveat here. In 2013, the MVP was given to LeBron James who happened to be playing the Power Forward position during his time with the Miami Heat. However, he is a traditional Small Forward. This means that the MVP award is more likely to won by smaller, faster, and versatile players than the traditional big men.

Lastly, upon looking at previous MVPs PER throughout their career, it can be seen that those who have won the award usually have an increase of two to four PER one or two seasons prior to winning the award with an exception of Kobe Bryant who won the award in 2008. It can also be seen that each player who wins the award are all at least above a PER of 22 one or two seasons prior. According to the benchmark, this means players need to be at least an All Star level player for at least one or two years to be consider an MVP. This doesn’t always happen but it’s rare for it not to.

**Win-Lose Ratio**

Although PER and Win Shares are valuable metrics to determine an MVP, Win Lose ratio should also be looked at since it reflects teams’ success. A player with high PER and WS doesn’t mean much if the player is on a losing team. In the past two decades, after plotting the Win Lose Ratio for each MVP’s team, majority of previous MVPs came from teams that had 70 – 75 winning percentage. Furthermore, for a player to be considered a MVP candidate, the team must win at least a 60 percent of their games. In the end, this percentage along with WS help quantify a player’s contributions to the team’s success.

**Statistical Inference**

For this part of the Capstone Project, I performed a hypothesis test to validate whether the true mean of the marginal Player Efficiency Rating (PER) between the Most Valuable Player (MVP) and the league average is 15. So far, only the PER threshold has been analyzed for MVP award winners, meaning what PER a player needs to have in order to be considered an MVP. However, since each season is different, certain seasons can have a lower or higher PER. Therefore, just strictly viewing the PER isn’t enough. To determine how well the MVP performed relatively to the league, the marginal PER can be analyzed. The marginal PER is calculated by taking the difference between the MVP PER to the league average for that season. In doing so, this value will allow us to determine how well a MVP candidate on average needs to perform relatively to the league for that season since each season is different.

The average marginal PER was determined through generating a sampling distribution of the sample mean. Using the data that I have, I used the bootstrap method to generate many independent random samples to compute the average marginal PER. After computing this measured statistic 10000 times, the mean of the sampling distribution of the sample mean was determined to be 15 and was used to estimate the true population mean.

To begin the hypothesis test, the null and alternative hypotheses were stated. The null hypothesis is that the average marginal PER equals 15 while the alternative hypothesis is the marginal PER does not equal 15. With the significance level was set to .05, if the calculated p value is lower than the significance level, then the null hypothesis will be rejected and the alternative hypothesis will be accepted. Vice versa, if the p value is higher than the significance level, then there isn’t enough evidence to refute the null hypothesis.

A one sample t test was used since the population variance was unknown. The t score was

calculated to be t = 0.46179937561216977 with a p value of p = 0.64767081755407196. This

means that assuming the null hypothesis is true, the probability of getting the sample mean from the random sample is 64%, which is higher than the set significance level. As a result, the

null hypothesis was not rejected. Lastly, the 95% confidence interval was calculated to be [14.21, 16.29] with this random sample.

As a result, this concludes that the usual MVP award winner on average has a higher PER of 15

compare to the league average.

**Conclusion**

In summary, for upcoming players to win the MVP award, these criteria are usually met. They are between the ages of 24-30. They were an All Star for at least two seasons. They come from teams that have won at least 60% percent of their games while majority of them win 70-75% of their games. Their PER and marginal PER are at least 23 is around 15 respectively.

**References**

“Calculating PER.” *Basketball-Reference.com*, www.basketball-reference.com/about/per.html.

Hollinger, John. “What Is PER?” *ESPN*, ESPN Internet Ventures, 8 Aug. 2011, www.espn.com/nba/columns/story?columnist=hollinger\_john&id=2850240.

“NBA Win Shares.” *Basketball-Reference.com*, www.basketball-reference.com/about/ws.html.

*What Criteria Matters Most in Making Your MVP Decision?* 12 Apr. 2017, 11:35 AM ET, www.nba.com/article/2017/04/12/blogtable-what-criteria-matters-most-making-mvp-decision#/.

Yang, Yuanhao. “Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics.” *University of California, Berkeley*, 2015.