The 2019-20 NBA Season: What Could Have Been

Ryan Kohanski Rithvik Saravanan Howard Yong

May 11, 2020

Abstract

The COVID-19 pandemic has affected our society in various ways and has changed numerous events on schedule for 2020. One such event that we were looking forward to was the end of the 2019-20 NBA season as well as the 2020 NBA Playoffs. Since a large portion of the 2019-20 NBA regular season games have already been played, we utilized the data collected from these games to run regression prediction models and calculate Elo ratings for each team in order to predict the standings for 2019-20 as well as the matchups and results of the Playoffs and the NBA season awards. Running our analysis by predicting the 7-game playoff series matchups, we predicted that the Western Conference Finals matchup will be between the #1 seed Los Angeles Lakers and the #2 seed Los Angeles Clippers and the Eastern Conference Finals matchup will be between the #1 seed Milwaukee Bucks and the #2 seed Toronto Raptors. [**NEED TO FIX THE MATCHUPS BASED ON BRACKET**] We found that running these simulations predicts the NBA Finals matchup between the Los Angeles Lakers and the Milwaukee Bucks with the Los Angeles Lakers ultimately claiming the Larry O'Brien NBA Championship Trophy. Additionally, we used several types of regression models to best predict end-of-season statistics for every player. We ultimately used a logistic regression model to predict the end-of-season statistics and leaders for each of the major categories and found that our predictions indicate that Giannis Antetokounmpo will claim both the NBA Most Valuable Player (MVP) award as well as the Defensive Player of the Year (DPOY) award. According to our prediction model, this will mark only the third time in NBA history that a player will win MVP and DPOY in the same season with the previous two players being basketball legends Michael Jordan and Hakeem Olajuwon.

Introduction

Due to the widespread impact of the COVID-19 pandemic throughout the world, almost every company, organization, and public event has canceled or suspended any activities that involve interpersonal contact for the forseeable future. Many of these activities are moving to a virtual format if possible, but several others have been forced to shut down.

As avid sports fans, the absence of the major sporting events during this time has hit us and many others around the world especially hard [1]. Some of the events that we particularly were looking forward to include the NBA, NCAA March Madness tournament, MLB, and the 2020 Summer Olympics.

In our curiosity, we decided to utilize this opportunity to exercise our data science and modeling skills in order to predict what could have been. Specifically, we focused on the NBA and the NBA Playoffs. Since the 2019-20 NBA season was suspended approximately one month prior to the end of the regular season (and the beginning of the Playoffs), we used the 2019-20 season data accumulated from the games played before the suspension to predict how the season and the Playoffs would have ended had everything gone according to schedule.

In this analysis, we will examine data from the 2019-20 NBA season as well as some data from previous NBA seasons in order to draw some meaningful conclusions about the remainder of the 2019-20 NBA season including the final season standings, playoff matchups, championship winner, and season award winners.

Methods

Predicting the 2019-20 NBA Season Standings

Since we missed one of the most exciting times of the year (the NBA Playoffs & Finals), we made some predictions on how the rest of the season might have played out using a popular methodology referred to as the Elo ratings system [2]. This tool, created by Hungarian-American physics professor Arpad Elo, was orginally designed to rate chess players, but is now used for all sorts of competitions ranging anywhere from sports to video games. This is a methodology that FiveThirtyEight and many other popular sports analysts take advantage of due to its simplicity and effectiveness [3].

These ratings depend only on the final score of each game as well as where it was played (home-court advantage). In other words, this system is built on a Win/Loss basis. We will be analyzing the 2018-19 NBA Season in its entirety to validate its performance, then we will apply it to the 2019-20 regular season in order to predict the matchups for the Playoffs and the Finals and ultimately the NBA Champions. For this project, we retrieved several types of data sources including game-by-game scores and schedules for several seasons from Basketball-Reference.com [4].

How does Elo work?

The long-run average for an Elo score in the NBA sits around 1500. An Elo of 1500 means that the teams performance would be normally distrubuted around an average of 1500 with the chance of performing better or worse. For more detail, Figure 3 (Appendix) shows what an Elo rating tells us about a team and how it can convey the teams overall season record. A higher Elo rating indicates that the team has a high win-loss ratio and is more likely to play deeper into the season.

The formula for Elo below shows how the probability of one team beating another is calculated using the ratings. When Player A competes in a match against Player B, Player A has an expected outcome (probability or score) for Team A ($E[A]$) where R_A is the rating for Team A and R_B is the rating for Team B. The expected outcome for Team A ($E[A]$) can be calculated by the formula below:

$$E[A] = \frac{1}{1 + 10^{\frac{(R_B - R_A)}{400}}}$$

The same calculation ($E[B]$) has to be done for Player B, but with R_A (current rating A) and R_B (current rating B) swapped so that $E[A] + E[B] = 1$. Once the match is played and S_A (actual outcome or score for Team A) and S_B (actual outcome or score for Team B) are determined, R'_A (the new rating for A) and R'_B (the new rating for A) are calculated with the formula below:

$$R'_A = R_A + K(S_A - E[A])$$

The S value in our case would either be 1 for a win, or 0 for a loss. This is because there are no ties in the NBA.

In this equation, K is an optimization constant that usually takes different values according the sport and the amount of games available. In other words, this value is the maximum amount by which a score can change in one match. If K is set too high, the ratings will jump around too much; if K is set too low, Elo will take too long to recognize important changes in team quality. Determining the right value of K is an entirely different and more complicated topic, so for this experiment we will be using $K = 20$, the optimal K for the NBA determined by FiveThirtyEight [3]. This is higher than most other sports and can likely be attributed to the fact that the NBA plays more games (81 games per team) and is subject to relatively little randomness.

Home-court advantage is set as equivalent to 100 Elo rating points. One hundred Elo points is equivalent to about 3.5 NBA points, so it can also be interpreted as the home team being favored by 3 to 4 points if the teams were otherwise evenly matched (obviously this value fluctuates from season to season). Since every team plays about half of their games at home and the other half away, a change in the home-court advantage value does not produce a significant difference in the ratings, but is still an important factor to consider.

Elo strikes a nice balance between ratings systems that account for margin of victory and those that do not. While teams always gain Elo points after wins and lose Elo points after losses, they also gain or lose more with larger margins of victory.

This works by assigning a multiplier to each game based on the final score and dividing it by a team's projected margin of victory conditional upon having won the game. For instance, the Golden State Warriors' 4-point margin over the Houston Rockets in Game 1 of the 2018-19 Western Conference finals was lower than Elo would expect for a Warriors win. So the Warriors gain Elo points, but not as many as if they'd won by a larger margin. The formula accounts for diminishing returns; going from a 5-point win to a 10-point win matters more than going from a 25-point win to a 30-point win. For the exact formula, see the footnotes.

Instead of resetting each team's rating when a new season begins, Elo carries over a portion of a team's rating from one season to the next. This is to account for any momentum that a team may build from season-to-season (i.e. sports dynasties). In NBA ratings, three-quarters of the previous score are kept. The high fraction reflects the fact that NBA teams are more consistent from year to year. For example, the Miami Heat ended the 2012-13 NBA season with an Elo rating of 1754. The team's Elo rating for the start of the 2013-14 season is calculated as follows:

$$(0.75 * 1754) + (0.25 * 1500) = 1692$$

Since this is a consistent method, we will also initialize the Elo scores for the 2019-20 NBA Season using the Elo scores from the 2018-19 season.

After incorporating a constant for home court advantage, our formula is as follows with $A = 100$ points (the value we previously determined represents a home-court advantage):

$$P(\text{Home team wins}) = \frac{1}{1 + 10^{-\frac{(H-R+A)}{400}}}$$

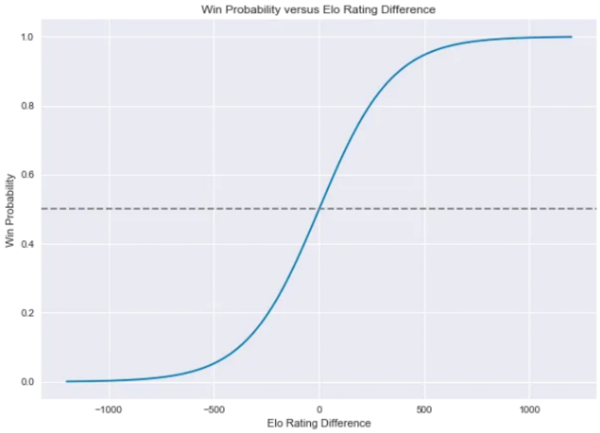


Figure 1: Logistic Function of Win Probability by Elo Rating Difference

In Figure 1, we see an example of a logistic function for win probability by Elo rating difference.

Predicting the 2019-20 NBA Season Award Winners

Another interesting part of any NBA season is the awards given to the players and teams based on their regular season performances. Some key awards that catch headlines every year include Most Valuable Player (MVP) and Defensive Player of the Year (DPOY). In addition to predicting the end-of-season standings, we decided that any analysis of the remainder of the 2019-20 NBA season would be incomplete unless it discussed award winners in the major categories. As avid basketball fans, this idea resonated with us so we decided to scientifically predict who would ultimately win the MVP and DPOY awards.

In order to predict award winners at the end of the season, we needed to predict the leaders of some of the crucial statistical categories at the end of the season. To predict these players, we analyzed some significant

Table 1: Elo Ratings for Every NBA Team (Descending)

	Team	Elo Rating
	Milwaukee Bucks	1628.4919
	Los Angeles Lakers	1627.8301
	Toronto Raptors	1601.2836
	Los Angeles Clippers	1588.4583
	Oklahoma City Thunder	1582.6036
	Boston Celtics	1569.6068
	Denver Nuggets	1560.4563
	Utah Jazz	1559.3314
	Houston Rockets	1546.1499
	Indiana Pacers	1544.9614
	Philadelphia 76ers	1536.7509
	Miami Heat	1531.0399
	Dallas Mavericks	1530.4395
	Memphis Grizzlies	1512.1054
	New Orleans Pelicans	1507.3374
16	Sacramento Kings	1496.7678
17	Brooklyn Nets	1491.0335
18	Orlando Magic	1490.3970
19	Portland Trail Blazers	1478.5000
20	San Antonio Spurs	1478.2600
21	Phoenix Suns	1453.5151
22	Washington Wizards	1446.2512
23	Charlotte Hornets	1435.6679
24	New York Knicks	1430.9475
25	Atlanta Hawks	1424.2177
26	Cleveland Cavaliers	1412.5029
27	Chicago Bulls	1410.1883
28	Minnesota Timberwolves	1391.4969
29	Golden State Warriors	1387.3738
30	Detroit Pistons	1383.5339

statistics and identified the optimal regression model to predict these statistics. Some of the statistics we utilized in building this model include true shooting percentage (TS%), total rebound percentage (TRB%), assist percentage (AST%), and block percentage (BLK%) among 22 total recorded categories. For additional information on the exact statistics that were used in these predictions, refer to Table 3 (Appendix).

Accordingly, we examined different types of regression models in order to identify which type of model best predicted some of the major statistics. These include win shares (WS), value over replacement player (VORP), player efficiency rating (PER), usage percentage (USG%), offensive box plus/minus (OBPM), and defensive box plus/minus (DBPM). For more detailed explanations of the significance of each of these statistics, refer to Table 3 (Appendix).

To identify the best prediction model, we first predicted WS from the current 2019-20 season statistics using 4 different regression models: linear, lasso, ridge, and logistic. We measured the performance of every model with the actual WS values for each player using RMSE (root mean squared error) to determine which had the least error where smaller RMSE values indicated higher accuracy. In order to reduce Monte Carlo variability, we used 200 repeated random samples of the data for each model to find the true RMSE values.

We then used this to predict the MVP and DPOY by looking at the leaders at the end of the season in WS, VORP, PER, USG%, OBPM, and DBPM because these categories carried significant weighting in determing the respective awards. In order to produce statistics that would reflect the end-of-season data, we updated each player's stats based on their team, position, schedule matchups, and usage percentage. We weighted each of these features by category and used the current player statistics to simulate the expected statistics for every player at the end of the 2019-20 season. We chose these specific features because a player's team and schedule can heavily influence their output, the position they play directly affects which stats are affected, and their usage percentage indicates how heavily their team relies on that specific player (which correlates to more playing time). For example, a point guard is more likely to focus on assists, a shooting guard is more likely to focus on shooting percentage and 3-point attempt percentage, and a center is more likely to focus on rebounds and blocks. Similarly, a player with a high usage percentage will be instrumental to the team and will thus receive more playing time to add to their stat lines.

Another important note to consider is that some of the statistical categories are related to other categories. For example, offensive win shares (OWS) and defensive win shares (DWS) are directly related to overall win shares (WS) because they are simply more specific aspects of the general WS category. In order to account for these confounding variables and ensure that the predictions were accurately estimated from all of the relevant data, we made sure to exclude the respective confounding variables when running each regression model. For instance, we excluded BPM when running regression models on OBPM and DBPM for the same reason.

Table 2: End-of-Season 2019-20 Predicted Stat Leaders with Logistic Regression Model

Player	Predicted WS	Player	Predicted VORP
Giannis Antetokounmpo	13.937733	Giannis Antetokounmpo	6.1011415
James Harden	13.140777	James Harden	5.5231850
Anthony Davis	11.683777	LeBron James	5.1576365
LeBron James	11.326877	Luka Dončić	4.9192570
Damian Lillard	11.172637	Anthony Davis	4.7997255

Player	Predicted PER	Player	Predicted USG%
Giannis Antetokounmpo	31.167695	James Harden	36.315868
James Harden	27.868148	Giannis Antetokounmpo	36.226468
Luka Dončić	27.691997	Damian Lillard	35.512591
Anthony Davis	27.423154	Luka Dončić	35.499674
Kawhi Leonard	27.183717	Bradley Beal	33.899727

Player	Predicted OBPM	Player	Predicted DBPM
James Harden	8.8347595	Giannis Antetokounmpo	3.6036553
Damian Lillard	8.4656587	Kris Dunn	3.0046294
Giannis Antetokounmpo	7.6219145	Anthony Davis	2.9304469
Luka Dončić	7.2672222	Nikola Jokić	2.6014332
LeBron James	7.1783205	Bam Adebayo	2.5128547

Results

[**Tables, figures, and text that illustrate your findings. Keep the focus on the numbers here. You will interpret your results in the next section.**]

First, we calculated the Elo ratings for each team for the games played so far in the 2019-20 season. As explained earlier, we incorporated 25% of the previous season's Elo ratings with 75% of this season's current Elo ratings. Table 1 shows the Elo ratings we derived for each team when the NBA season was suspended. To better visualize how each team's Elo rating is updated throughout the season, Figure ... shows the Elo rating of every team since the opening of the 2019-20 season for all games that have been played (prior to the suspension). ... [**NEED TO WRITE ABOUT ELO PREDICTIONS AND PLAYOFFS/FINALS HERE**]

In addition to simulating how the 2019-20 NBA season and Playoffs would have ended, we tested several regression models to determine the best model to predict end-of-season WS. By testing linear, lasso, ridge, and logistic regression models, found that the lasso model provided the lowest RMSE value (0.24835944) with the linear model (0.34105176) and logistic models (0.33794665) providing slightly higher RMSE values.

We then predicted the season leaders for the various principal categories of WS, VORP, PER, OBPM, and DBPM. We utilized a logistic regression model for predicting the stat leaders due to its relative simplicity and clarity. We verified that this logistic regression model was appropriate by measuring its RMSE values and plotting predicted end-of-season statistics for each player for every category as well as the actual vs. predicted statistics for each category for the current 2019-20 season data (Figure 2). The plots in Figure 2 indicate the end-of-season statistic value for each category based on a player's current 2019-20 season data for that statistic. The red lines indicate the value that a player can expect to end the season with for every value along the x-axis (which represents the player's current statistic for that category). These plots verify that the predicted stats fall within a very small margin of the actual stats and thus have small residual values (since a majority of this season's games have already been played). Also, there is a very minor shift in some of the data points (i.e. VORP) because the last month of games would likely change some of these player statistics since teams with guaranteed playoff seeding are more likely to rest their star players.

Using these accumulated season statistics, we determined the end-of-season statistics by updating each players' stat lines based on team, position, schedule matchups, and usage percentage stats as explained in detail in the Methods section.

The final season stat leaders for each of the categories is shown in Table 2. We verified these predictions by limiting the qualifiers for each category and comparing with each player's previous performance history. We narrowed the pool of players down by only considering players with more than 25 games played and 1200 minutes played because this reflects the criteria that the NBA Season Awards use to nominate qualifying players. By inspecting the top five players in each category, we noticed several household names and early season favorites for MVP and DPOY including Giannis Antetokounmpo, James Harden, LeBron James, and Anthony Davis.

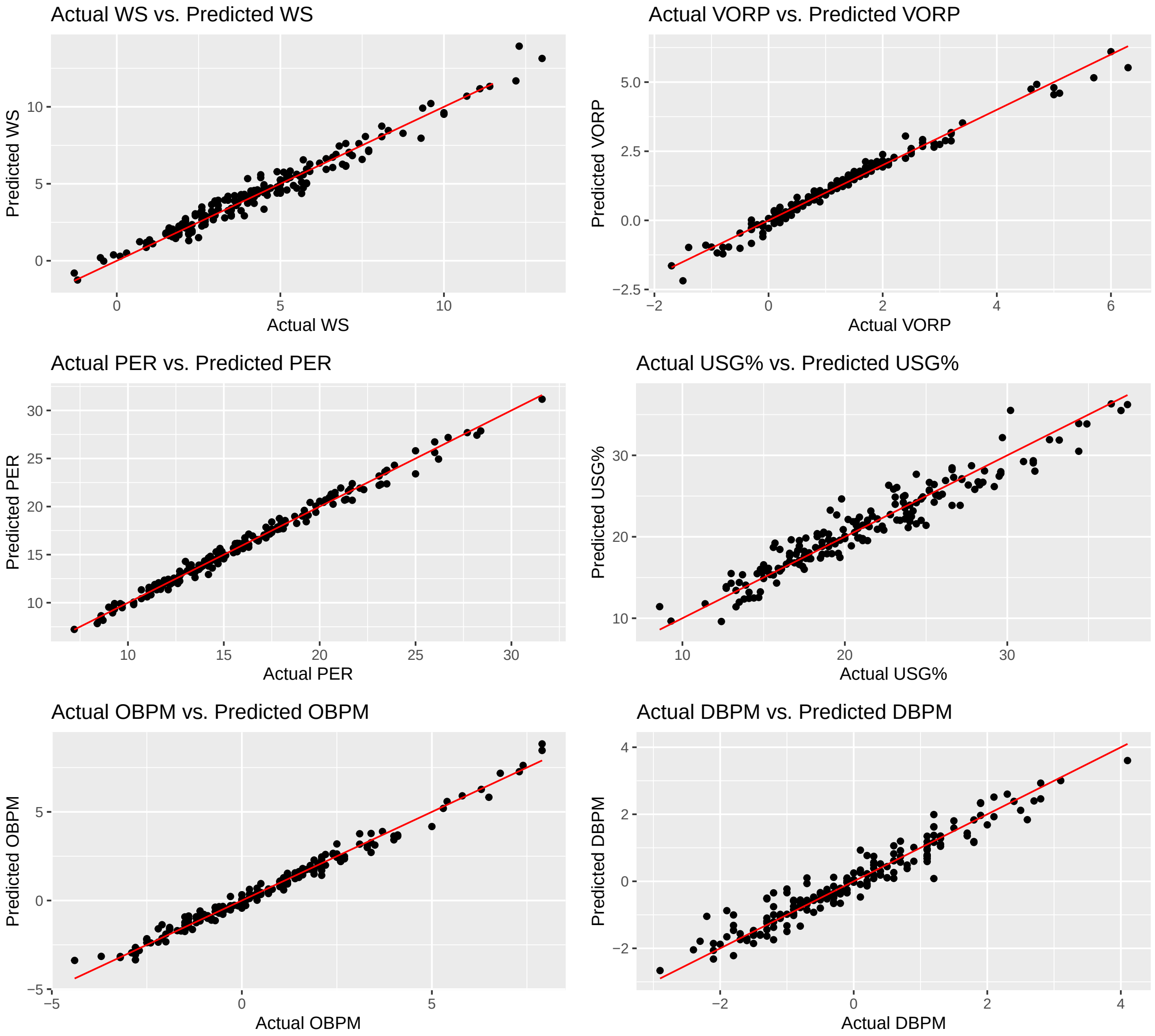


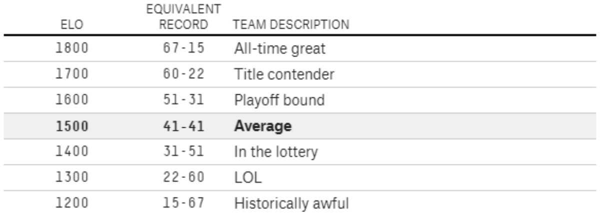
Figure 2: Predictions of Key Statistics with Logistic Regression Model

Conclusion

[Interpret what you found. What are the main lessons we should take away from your report?]

Shortcomings/Drawbacks

Appendix



ELO	EQUIVALENT RECORD	TEAM DESCRIPTION
1800	67-15	All-time great
1700	60-22	Title contender
1600	51-31	Playoff bound
1500	**41-41**	**Average**
1400	31-51	In the lottery
1300	22-60	LOL
1200	15-67	Historically awful

Figure 3: Breakdown of Elo Rating

Statistics	Meaning
G	Num. games played
MP	Num. minutes played
PER	Measure of per-minute production
TS.	Overall shooting efficiency
X3PAr	% of field goal attempts from 3-point range
FTr	Num. free throw attempts per field goal attempt
ORB.	% of available offensive rebounds that a player grabbed
DRB.	% of available defensive rebounds that a player grabbed
TRB.	% of available total rebounds that a player grabbed
AST.	% of teammate field goals that a player assisted
STL.	% of opponent possessions that were stolen by a player
BLK.	% of opponent field goals attempts that were blocked by a player
TOV.	Num. turnovers committed per 100 plays
USG.	% of team plays used by a player
OWS	Num. wins contributed by a player from his offense
DWS	Num. wins contributed by a player from his defense
WS	Num. wins contributed by a player
WS.48	Num. wins contributed by a player per 48 minutes
OBPM	Offensive points per 100 possessions above a league-average player
DBPM	Defensive points per 100 possessions above a league-average player
BPM	Total points per 100 possessions above a league-average player
VORP	Points per 100 team possessions contributed by a player above a replacement-level player

References

[1] List of all sporting events canceled around the world during the coronavirus pandemic (https://www.espn.com/olympics/story/_/id/28824781/list-sporting-events-canceled-coronavirus)

[2] Elo ratings system (https://en.wikipedia.org/wiki/Elo_rating_system)

[3] FiveThirtyEight NBA Elo Ratings (https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/)

[4] Compilation of in-depth NBA statistics (https://www.basketball-reference.com/)

[5] Elo Ratings for NBA Teams (http://practicallypredictable.com/2018/04/15/elo-ratings-for-nba-teams/#more-1019)