Predicting the Probability of Being Selected as NBA All-star Rosters based on NBA Players' Statistics using Logistic Regression Models

Yuanjie Ji 1004768469

2020/12/22

Github repository: https://github.com/Yuanjie-Ji/STA304-0101-Final-Yuanjie-Ji/

Keywords: Statistics, Logistic Regression, Basketball, NBA All-Star

Abstract

The objective of this research is to build models that use NBA players' per game statistics to predict NBA All-Stars. The NBA players' per game statistics data from 2015 to 2019 are downloaded and processed to build logistic regression models for each position on the court. Then the 2020 per game statistics data are used as the test data to examine the accuracy of the model. The 2021 All-Star list can be predicted by fitting 2021 per game statistics into the models in the near future.

Introduction

NBA, National Basketball Association, is a professional basketball league that represents the highest level of men basketball playing. Entering the NBA means becoming one of the best basketball players in the world. In the middle of each season of NBA, a special events called NBA All-Star will be held. Top 30 players will be selected to play an exhibition game based on the results of voting by fans and coaches.

NBA All-star has long been considered an exceptional honor to NBA players. Being an NBA All-star player means being one of the best current NBA players. One can become an NBA All-star by having great performances in games, therefore winning wide appreciation from NBA coaches, players, reporters, and NBA fans. Being an NBA All-star not only means being one of the top 30 players among more than 400 players in the NBA league but also can have huge positive impact on one player's salary, business value, trading value, and chance pf entering Naismith Basketball Hall of Fame.

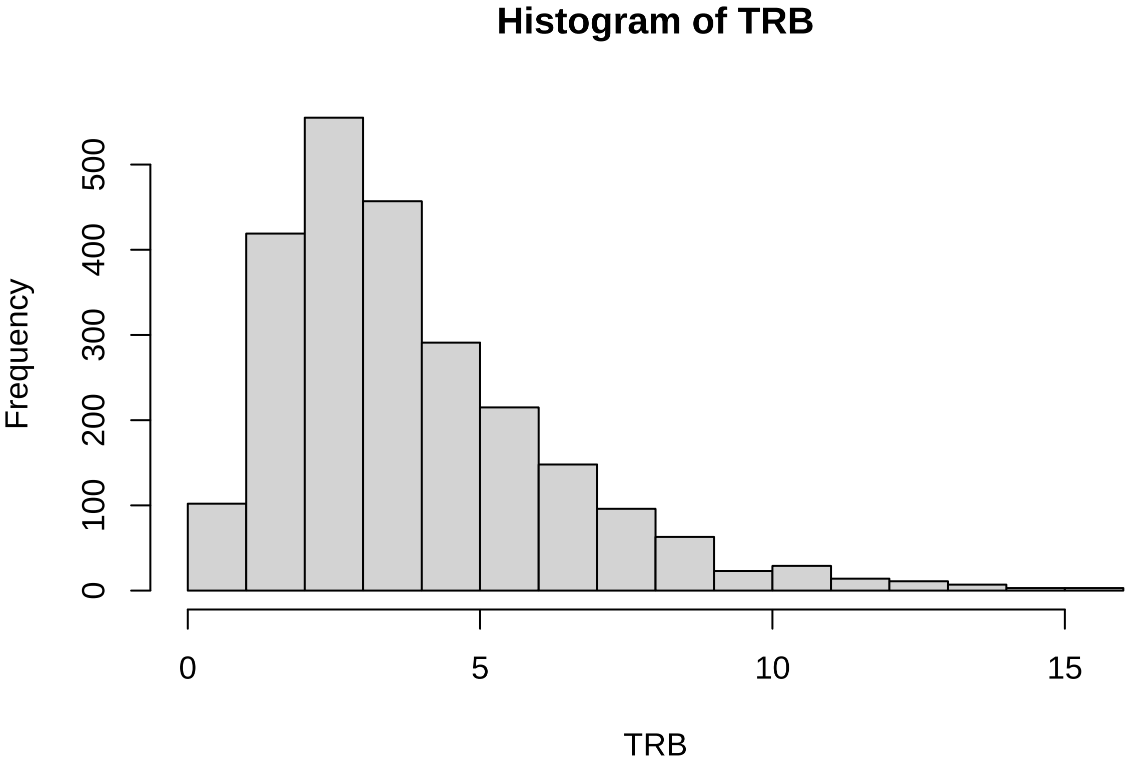
Due to such importance of NBA All-star, NBA fans are never tired of discussing which group of players can become the NBA All-stars. The goal of this research is to analyze the relationship between the probability of being an NBA All-star and NBA players' statistics using a logistic regression model. In the model, a binary variable of being an NBA All-star or not is constructed and is used as the response variable. Since per game statistics are the most direct statistics that reflect players' performances, the players' per game statistics, including points, rebounds, assists, turnovers, steals and blocks, are selected as the predictors. Since the 2020-21 season has not started, the data of 2019-20 season are fitted into the models, and a simulated list of All-Star players will be produced. The simulated list will be compared with the real list to see the accuracy of the models. The models can be used to predict 2021 NBA All-Stars when there are enough data for 2020-21 season.

Data

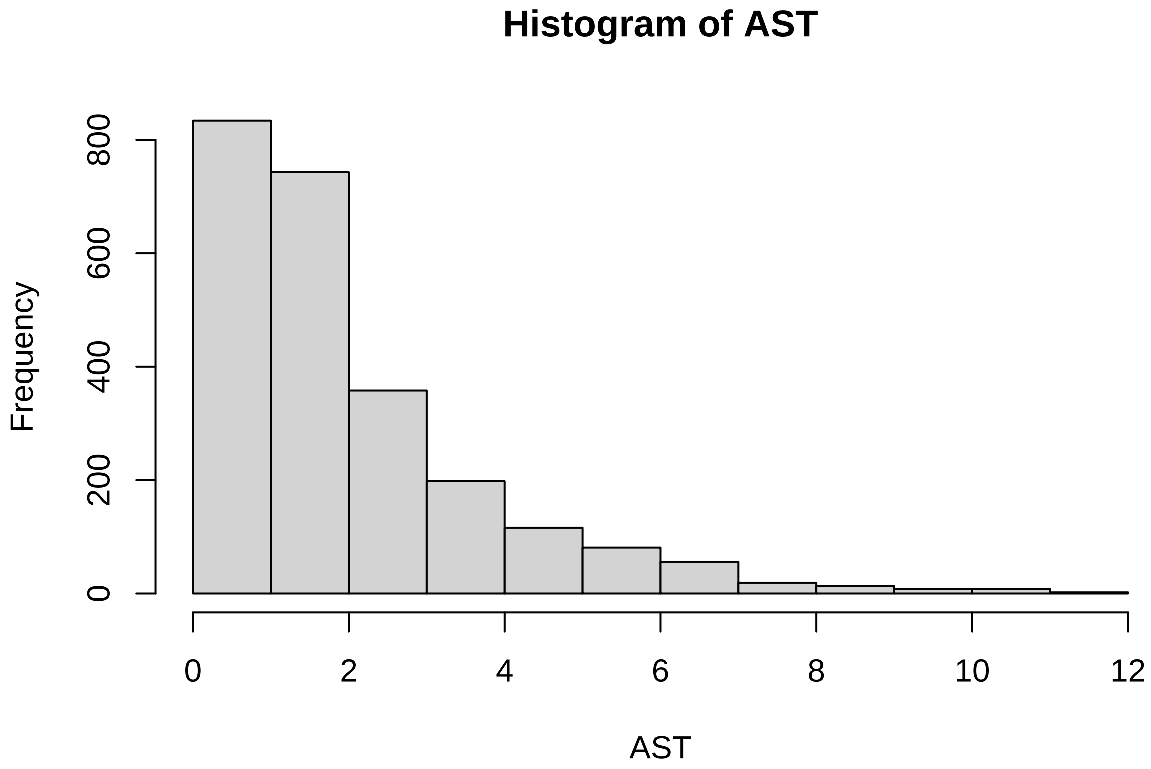
The data used in the research are downloaded from Basketball Reference 2014-15 NBA Player Stats: Per Game (Corby, 2015), Basketball Reference 2015-16 NBA Player Stats: Per Game (Corby, 2016), Basketball Reference 2016-17 NBA Player Stats: Per Game (Corby, 2017), Basketball Reference 2017-18 NBA Player Stats: Per Game (Corby, 2018), Basketball Reference 2018-19 NBA Player Stats: Per Game (Corby, 2019), and Basketball Reference 2019-20 NBA Player Stats: Per Game (Corby, 2020).

The following columns of data are selected as the explanatory variables:

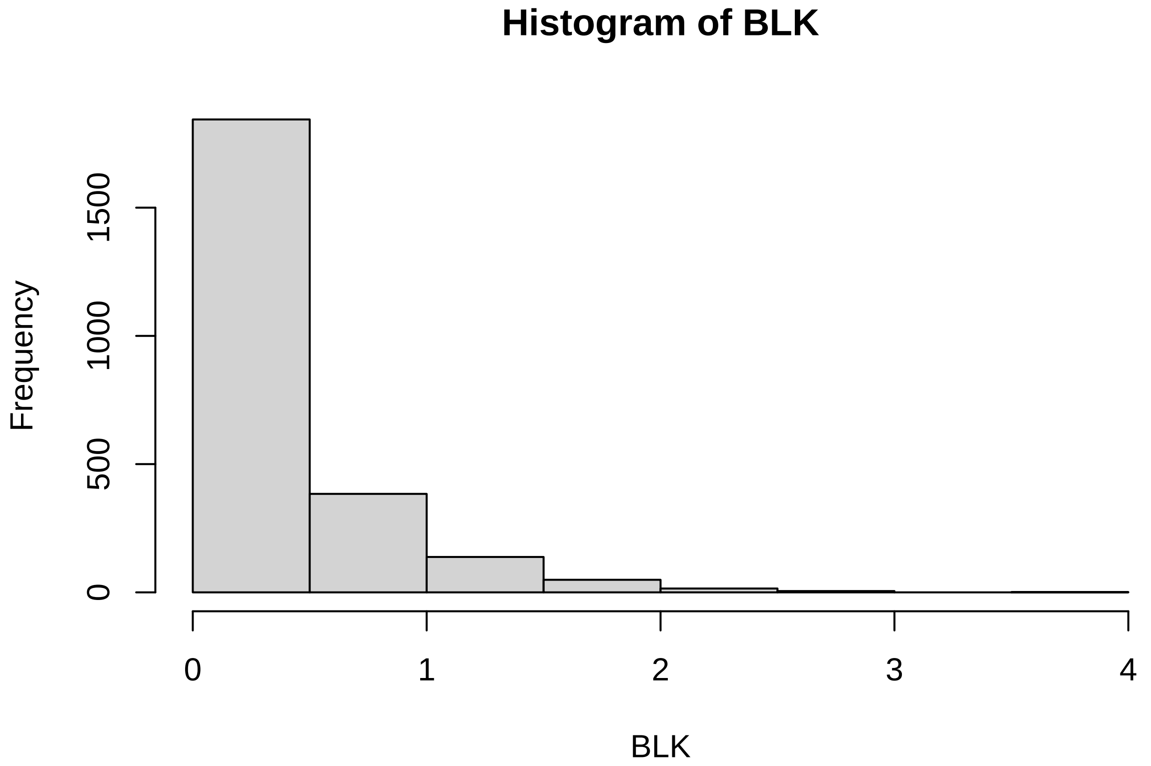
TRB: the sum of numbers of offensive rebounds and defensive rebounds per game of a player. Continuous variable.



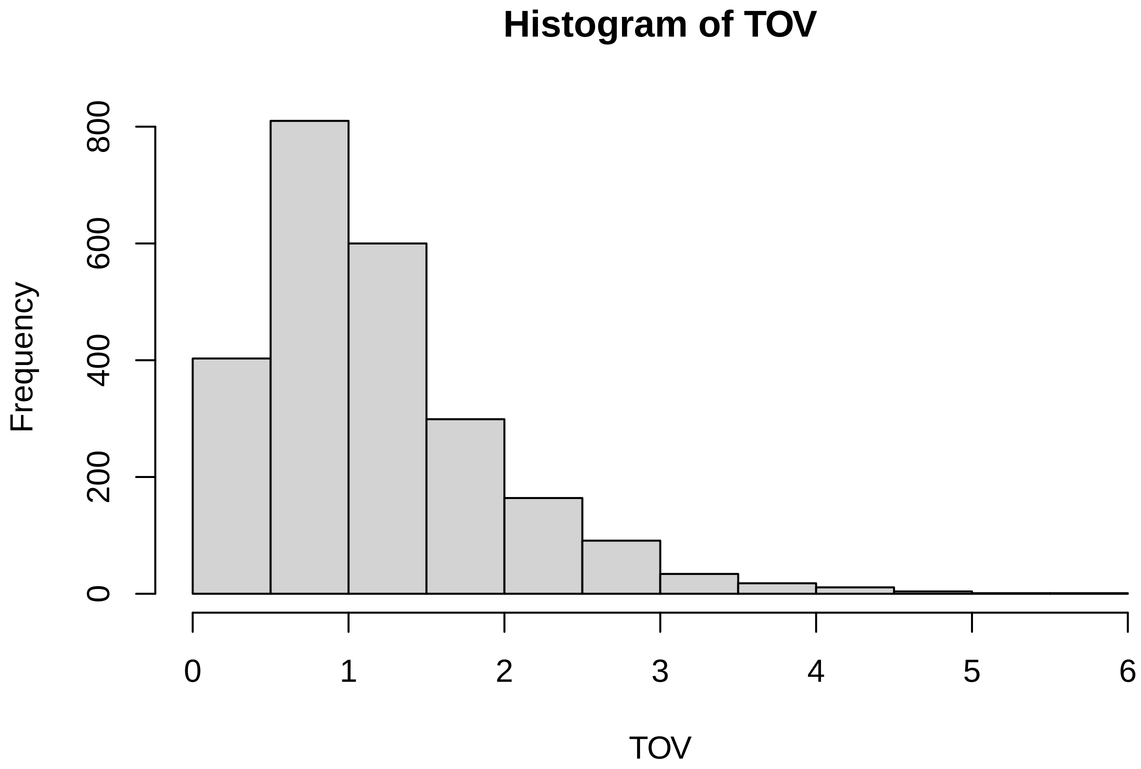
AST: the number of assists per game of a player. Continuous variable.



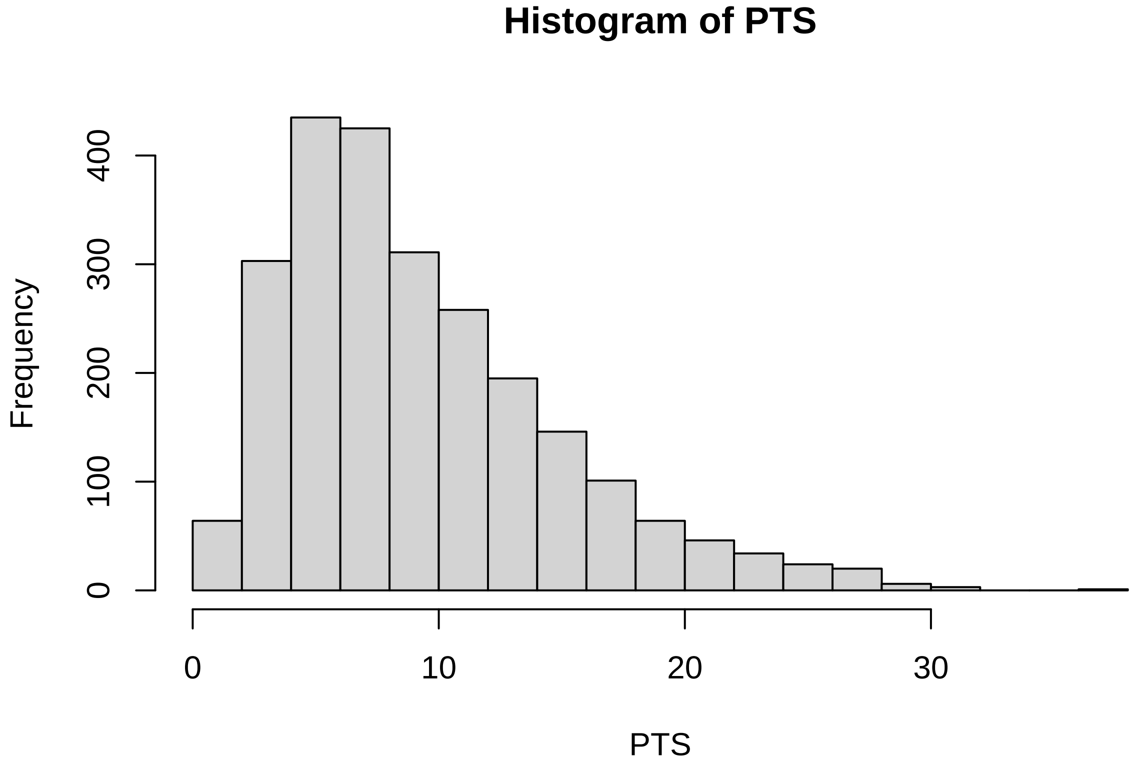
BLK: the number of blocks per game of a player. Continuous variable.



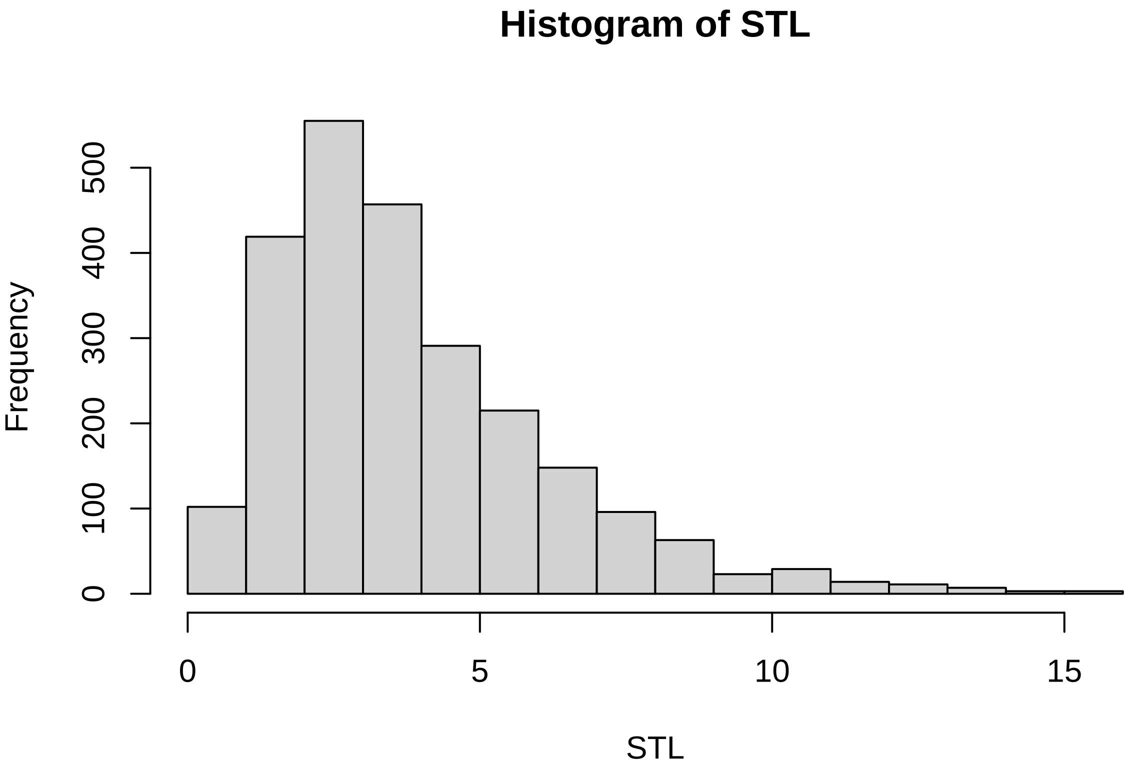
TOV: the number of turnovers per game of a player. Continuous variable.



PTS: the number of points per game of a player. Continuous variable.



STL: the number of steals per game of a player. Continuous variable.



Then a binary variable, ALL.STAR, is manually added to the data set according to the All-Star lists from 14-15 season to 19-20 season. The value 1 represents the player is selected as one of the All-Star rosters, and 0 represent the player is not selected. This variable will be later used as the response variable of the models.

Players who play under 20 games are removed from the data set. This is because the data of a player who play too few games are often biased. For example, a player at the end of team rotation got a chance to play and played well in a game when the usual starters are rested by the coach, and he never got a second chance to play during the whole season. Then his per game statistics would seem great despite the fact that he was not a good player. Another main reason of too few games played is injury. Since injured players would also not be considered being selected as All-Star players (unless get injured after being selected), a threshold of 20 minimum games is used to reduce such cases.

Model

Since the response variable selected in this research is a binary variable, a logistic regression model is suitable for this research. However, there are multiple positions in basketball, and each position have different responsibilities on the court. Therefore, the performance of players of different positions should be reflected by different response variables. Point guards are responsible for organizing the team's offense by controlling the ball and passing the ball, so assists are more important for point guards than for other positions. Since point guards are usually the faster but shorter players in the team, it can be predicted that blocks and rebounds will not be a significant predictor for point guards. Shooting guards, as the name goes, mainly create long-ranged shots on the court. As a result, points will be the most important predictor for shooting guards. Shooting guards are still usually shorter than players of other positions beside point guard, so blocks predictor may also not be significant. Small forward is a all-around position: small forward players can shoot or break through and layup on the offense side and steal the ball and grab rebounds sometimes on the defense side. Consequently, all predictors may be significant for small forward players. Compared with small forwards, power forwards need to score near the basket, though modern power forwards also shoot long ranged shots. Moreover, power forwards also do some defensive job, including grabbing rebounds and blocking shots. Protecting the rim and grabbing rebounds are the two main tasks for centers. Some centers are also good at scoring, so it is assumed that rebounds, blocks, and points are significant to centers.

Confronting such differences between positions, it is necessary to construct multiple logistic regression models , and the models will be constructed using R (R Core Team, 2019). Each model will only include the players of the same position. According to the NBA All-Star rosters voting rules (), players' positions are only divided into two categories: guards (point guard and shooting guard) and frontcourt players (small forward, power forward, and center). Then I will construct 3 logistic regression model accordingly: one for guards, one for small and power forward, and one for center. Center is separated from the frontcourt because there must be one person responsible for the center's tasks on the court. Also, a interesting trend is found from the model for center and is discussed in the "Results" part.

The format of the model for each position is:

$$log\ (\frac{p}{1-p}) = \beta_0 + \beta_1 X_{Age} + \beta_2 X_{TRB} + \beta_3 X_{AST} + \beta_4 X_{BLK} + \beta_5 X_{TOV} + \beta_6 X_{PTS} + \epsilon$$

p is the probability for a NBA player being selected as an All-Star player.
β_0 is the intercept term.
β_1 is the coefficient for the slope of X_{TRB}.
X_{TRB} represents the sum of offensive rebounds and defensive rebounds per game of each player.
β_2 is the coefficient for the slope of X_{AST}.
X_{AST} is the average number of assists per game of each player.
β_3 is the coefficient for the slope of X_{BLK}.
X_{BLK} refers to the blocks per game made by a player.
β_4 is the coefficient for the slope of X_{TOV}, which is number of turnovers per game made by a player.
β_5 is the coefficient for the slope of X_{PTS}, which is number of points scored by each player per game.
ϵ is the error term.

Results

Table 1: The Logistic Regression Model for Guards

para_G	coeffi_G	pvalue_G
Intercept	-0.1804	< 2e-16
TRB	0.0023	0.77145
AST	0.0301	5.23e-06
BLK	0.0782	0.05857
TOV	-0.0678	0.00041
PTS	0.0231	< 2e-16
STL	-0.0064	0.78733

The logistic regression model for guards matches the assumptions that All-Star guards need to organize the offense and score well. By looking at the p-values, only three predictors, assists, turnovers, and points, are significant for guards, and the coefficient of turnover is negative. This means that guards need to create offense opportunities (assisting other players scoring) and try to avoid turnovers by the same time. Scoring is also a main task for guards, especially for shooting guards.

Table 2: The Logistic Regression Model for Forwards

para_F	coeffi_F	pvalue_F
Intercept	-0.1229	0.0000

para_F	coeffi_F	pvalue_F
TRB	-0.0142	0.0021
AST	0.0767	0.0000
BLK	0.1127	0.0000
TOV	-0.0529	0.0164
PTS	0.0177	0.0000
STL	-0.0486	0.0093

The logistic regression model for forwards implies that All-Star forwards should be all-around players as assumed. Points, assists, blocks, and steals are all significant predictors with positive coefficients, suggesting that All-Star forwards need to play important roles in both offense and defense. They need to score by attacking the rim or shooting and defend the opponents by stealing the ball and blocking shots.

Table 3: The Logistic Regression Model for Centers

para_C	coeffi_C	pvalue_C
Intercept	-0.1528	0.0000
TRB	-0.0107	0.0460
AST	0.0528	0.0001
BLK	0.0041	0.8500
TOV	-0.0294	0.3090
PTS	0.0275	0.0000
STL	0.0035	0.9260

The logistic regression model for centers seems strange. Traditionally, centers are assumed to protect the rim and grab rebounds. However, in this model, points and assists are the only two significant predictors with positive slopes (blocks are not significant). Furthermore, although number of total rebounds is a significant explanatory variable, the coefficient of it is negative, which is contrary to the assumptions that All-Star centers would be good at grabbing rebounds.

Table 4: The Simulation Results

players	log_odds	probability
Bam Adebayo	0.3256023	67.9%
Giannis Antetokounmpo	0.4758438	74.9%
Bradley Beal	0.4446632	73.6%
Devin Booker	0.3405617	68.7%
Jimmy Butler	0.4583130	74.2%
Anthony Davis	0.5119611	76.5%
DeMar DeRozan	0.4423953	73.5%
Luka Doncic	0.4252548	72.7%
Joel Embiid	0.3510868	69.2%
James Harden	0.5002474	76.0%
Brandon Ingram	0.3708986	70.1%
LeBron James	0.4122211	72.1%
Nikola Jokic	0.4619876	74.3%
Kawhi Leonard	0.4824176	75.2%
Damian Lillard	0.5216172	76.9%
CJ McCollum	0.3100500	67.1%
Khris Middleton	0.3029799	66.8%
Donovan Mitchell	0.2891331	66.1%

players	log_odds	probability
Pascal Siakam	0.3440203	68.8%
Jayson Tatum	0.3302659	68.1%
Karl-Anthony Towns	0.5220175	76.9%
Nikola Vucevic	0.2891061	66.1%
Russell Westbrook	0.3228206	67.8%
Andrew Wiggins	0.3878224	71.0%
Trae Young	0.4252727	72.7%

The players with top 25 probability of being selected as All-Stars are listed above. This list can be deemed as the list of 2020 All-Star list simulated by the models. In this list, Andrew Wiggins, DeMar DeRozan, Karl-Anthony Towns, CJ McCollum, Bradley Beal, and Nikola Vucevic were not selected in reality while Kemba Walker, Kyle Lowry, Ben Simmons, Domantas Sabonis, Chris Paul, and Rudy Gobert were actually selected but not in the list simulated (6 among 25 are different). The most differences between the players in the simulated list and the players in the actual list is that players in the actual list are spiritual leaders and experienced while the players in the list have better statistics. These experienced leaders are respected by players, reporters, and fans, and usually have many positive influences that cannot be reflected by the statistics on the team. Some more interpretation about the differences between the simulated list and the real list is discussed in the "Weakness" part.

Discussion

Summary

In this research, multiple logistic regression models are constructed to predict NBA All-Stars. First, players' per game statistics from 14-15 to 19-20 seasons are downloaded from Basketball-Reference.com (Basketball-Reference.com, 2015) (Basketball-Reference.com, 2016) (Basketball-Reference.com, 2017) (Basketball-Reference.com, 2018) (Basketball-Reference.com, 2019) (Basketball-Reference.com, 2020). Data of 14-15 to 18-19 seasons are used to constructed the models, and 19-20 season data are used to test the models. Players are divided into three groups according to their positions on the court: guards, forwards, and centers. Then a logistic regression model is constructed for each group. Subsequently, the data of 19-20 season are fitted into the models. The list of All-Stars generated by the models are compared with the real 2020 All-Star list (NBA.com, 2020) to see if there are any differences. Finally, the interpretation of weird predictors, weakness of the models, and next steps are discussed.

Intepreting the Models

The coefficients of all the predictors for guards and all the predictors except rebounds for forwards corresponds to the assumptions made before constructing the models pretty well. The interpretation of these predictors has already been discussed in the "Results" part. This part will mainly discuss the weird things in the model.

One strange thing in the model is that the rebound explanatory variable is not a positive factor in any model. For guards, it is quite easy to explain: guards are usually not tall, and it is not their task to play near the rim. The negative slope of rebounds for forwards may be due to the nature of the data. There are quite number of forward players playing like centers do but are not tall enough to play as centers, so they play other positions but still do things that centers do. Some examples of this kind of players are Montrezl Harrell, Larry Nance Jr., and Tristan Thompson. They grab many rebounds per game but are mostly not selected as All-Star players. Since the All-Star forwards usually grab less rebounds then they do, the coefficient for the rebounds predictor become slightly negative.

It is a different story when coming to centers. In the model for centers, the rebounds predictor has negative slope the and the blocks predictor is not significant. These abnormal results can be explained by the evolution of modern centers. As the article “How the big men of the NBA adjusted to survive in the modern game” says, great modern centers are capable of dribbling, organizing the offense, and shooting long-ranged shots (HOOPSBEAST, 2019). The recent All-Star centers, like Nikola Jokic, Bam Adebayo, and Joel Embiid, have shown high level of skills in playing like guards do and do win more support than the traditional centers. Since the traditional centers focus more on rebounding and blocking shots, traditional centers’ absences in the recent All-Star games result in the negative coefficients of rebounds and the insignificance of blocks predictor.

Conclusion

By testing the models using 19-20 season data, the simulation made by the model is approximately accurate. The difference between the simulated All-Star list and real rosters list is not huge. Even for the players in the simulation list but not in the real list, they are still good enough to be selected. If the factors beyond players’ performances on the court were not taken into consider, this model would be quite suitable for predicting 2021 NBA All-Star rosters.

Weakness

One weakness of the models is that, while players’ levels of offense can generally be represented assists and points, players’ levels of defense are not well reflected by simply looking at steals and blocks. An important aspect of defense is limiting the opponents’ spaces to score and field goal rates. Ben Simmons is one of the best defenders among point guards. He is the tallest point guard (2.08m) and is faster than most of the point guards. Other point guards can seldom find opportunities to shoot or layup facing Simmons’ defense. His another advantage of defense is that he can defend opponents of any positions due to his perfect athleticism, so that he is good at defending “pick and roll” and “mismatch”. NBA coaches appreciate Ben Simmons’ defensive abilities and selected him into the All-Star games in 2019 and 2020. However, it seems that Ben Simmons is not favored by the models, since his abilities in offense are not prominent compared with other All-Star guards and his abilities in defense are not well reflected in the model.

Another weakness of the models is that the models may favor players with good offensive statistics but cannot help their teams win games much. Karl-Anthony Towns is one of the best players according to the result of simulation. However, his team has bad win rates, and he is blamed for bad defense and shooting too much. Since Karl-Anthony Towns was not helping his team win games, he did not receive many votes for 2020 All-Star.

A third weakness of the models is that the spiritual factors are overlooked by the models. Some players with great leadership and mental toughness are selected into All-Star games in reality, such as Chris Paul and Kyle Lowry. These players’ statistics may not look so good, but they are actually helping the team much. On the contrary, some playes favored by the models are criticized for bad attitudes despite their good-looking statistics. Jimmy Butler once insulted Karl-Anthony Towns and Andrew Wiggins by saying they are “soft” (Conway, 2018). These two players are in the simulated list but were actually not selected as 2020 All-Star rosters.

The last weakness of the models is that the models omit the factor of the size of fan groups. Bradley Beal and CJ McCollum had a perfect season in 19-20 season and are also in the simulated All-Star list. Nevertheless, the numbers of their fans are not as large as those of 2020 All-Star players. Unfortunately, they were not selected although their performances were great.

Next Steps

This research is designed to predict NBA 2021 All-Stars. However, since the 20-21 season has not started yet, the 19-20 season statistics is temporarily used to test the model. One next step is to wait for the season to start and fit the 20-21 season statistics into the model after more than 20 games played. Then the 2021 All-Stars can be predicted by the model.

Another future step is to include some more predictors into the model, such as the win rate of the player's team, the number of player's fans, and the player's defensive efficiency. These new predictors will help solve the problems mentioned in the "Weakness" part.

References

Corby, David. (2015). "2014-15 NBA Player Stats: Per Game". www.basketball-reference.com/leagues/NBA_2015_per_game.

Corby, David. (2016). "2015-26 NBA Player Stats: Per Game". www.basketball-reference.com/leagues/NBA_2016_per_game.

Corby, David. (2017). "2016-17 NBA Player Stats: Per Game". www.basketball-reference.com/leagues/NBA_2017_per_game.

Corby, David. (2018). "2017-18 NBA Player Stats: Per Game". www.basketball-reference.com/leagues/NBA_2018_per_game.

Corby, David. (2019). "2018-19 NBA Player Stats: Per Game". www.basketball-reference.com/leagues/NBA_2019_per_game.

Corby, David. (2020). "2019-20 NBA Player Stats: Per Game". www.basketball-reference.com/leagues/NBA_2020_per_game.

HOOPSBEAST. 2019. "How the big man of the NBA adjusted to survive in the modern game". https://www.hoopsbeast.com/evolution-of-the-nba-centers/.

Conway, Tyler. (2018). "Jimmy Butler Reportedly Insulted Karl Anthony-Towns, Andrew Wiggins: They Soft!". https://bleacherreport.com/articles/2800185-jimmy-butler-reportedly-insulted-karl-anthony-towns-andrew-wiggins-they-soft.

R Core Team (2019b). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.