2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress

Optimizing NBA Player Selection Strategies Based on Salary and Statistics Analysis

Ramya Nagarajan
Department of Computer Science
Prairie View A&M University
Prairie View, Texas 77446
Email: rnagarajan@student.pvamu.edu

Lin Li
Department of Computer Science
Prairie View A&M University
Prairie View, Texas 77446
Email: lilin@pvamu.edu

***Abstract*—In National Basketball Association (NBA), how to comprehensively measure a player's efficiency and how to sign talented players with reasonable contracts are two challenging issues. This research explored the key indicators widely used to measure player efficiency and team performance. Through data analysis, two indicators, namely Player Efficiency Rating and Player Defense Rating, were chosen to formulate the prediction of the team winning rate. Besides, different player selection strategies were proposed according to different objectives and constraints. Experimental results show that the developed team winning rate prediction models have high accuracy and the player selection strategies are effective.**

1. Introduction

In recent years, with the fast development of big data and business intelligence technologies, sports analytics gained increased attention. National Basketball Association (NBA), the premier men's professional basketball league in the world, is changing quickly with the rise of big data analytics [1]. A lot of research has been conducted to help athletes, coaches, and general managers to gain competitive advantage and commercial benefits. For example, Kubatko et al. [2] compared existing player performance measurement formulas and provided a general method to study the relationship between basketball possessions and the player's statistics. Piette et al. [3] investigated how individual player's performance was affected by teammates and proposed a network-based algorithm to estimate the statistical significances. Garcia et al. [4] studied different performance indicators which discriminate the teams' win-loss results in regular season and playoff games. Sadiq and Zhao [5] explored how to form a team by selecting players to optimize court coverage.

Generally, sports analytics can be categorized as on-field and off-field, with the former dealing with on-field performance improvement and the latter focusing on business improvement and profitability. Compared to the tons of research being conducted in analyzing NBA's on-field statistics [6], [7], [8], [9], [10], research on off-field, particularly, formulating the relationship between player performance and various business constraints to optimize business strategies still need to be explored.

To control costs and prevent competitive unbalance, starting from 1984-85 season, NBA annually sets a salary cap which limits the amount of the money a team can pay for his players. Heavy luxury tax can be charged if a team's payroll exceeds the cap. For example, the salary cap of the 2016-17 season is $94.143 million [11]. Due to the salary cap, a team's general manager must carefully evaluate the players' performance before offering contracts. A bad contract may lead to a team's loss of competitiveness and financial flexibility. How to effectively sign players from the market with limited budget is therefore the key to a franchise's business success. Despite its importance, there were many unsuccessful contract cases which, sometimes, even caused NBA to approve special amnesty provisions to alleviate the problems (e.g., the 2005 "Allen Houston Rule" and the 2010 Collective Bargaining Agreement). In view of this, people can't help asking *"Are the NBA general managers smart or not?"* Correspondingly, this paper aims to explore the best player selection strategies by studying the player statistics, team performance, and the salary cap.

The rest of the paper is organized as follows: Section 2 explains basic NBA statistics concepts and how the players' efficiency is measured and their relationship with the team performance. Section 3 describes our player selection strategies designed for different objectives and constraints. Section 4 presents the experimental results and analyses. Finally, Section 5 concludes the paper and proposes some future work.

2. Concepts and Preliminary Analysis

In NBA games, a player's efficiency is measured by a number of basketball statistics, including points, rebounds, assists, steals, blocks, turnovers, etc. Table 1 presents a list of notations representing the most commonly referred statistics. Some of the statistics show the positive contribution of the player (e.g., points and rebounds), and some reflect the negative contribution (e.g., turnovers). Some statistics measure the player's offense skills (e.g., assists), and some measure the player's defense skills (e.g., blocks and steals). As each statistics only shows partial contribution of the player, people always desire to coin an indicator to best gauge the player's overall strength.

978-1-5386-1956-8/17 $31.00 © 2017 IEEE
DOI 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.175

TABLE 1. PLAYER STATISTICS NOTATIONS

Notation	Meaning
GP	Games played
Min	Minutes played
PACE	Possessions in a game
PTS	Points
AST	Assists
REB	Total rebounds
OREB	Offense Rebounds
DREB	Defense Rebounds
STL	Steals
BLK	Blocks
FGM	Field Goal Made
FGA	Field Goal Attempted
3P	Three-point Field Goal Made
3PA	Three-point Field Goal Made
FTM	Free Throw Made
FTA	Free Throw Attempted
TO	Turnovers
PF	Personal Fouls

2.1. Overall Player Efficiency Measurements

In the past decades, sports analysts and statisticians developed different methods using the combination of individual statistics to measure a player's comprehensive performance. For example, the NBA's *Efficiency Rating* (EFF) [12] is a single number measure of the player's overall contribution (both positive and negative) to a game he plays in. The formula is: (PTS + REB + AST + STL + BLK - ((FGA - FGM) + (FTA - FTM) + TO)) / GP. EFF is simple, but it doesn't consider the player's contribution in terms of his team's playing styles, and it doesn't reflect the player's role as a starter or a bench player. Plus, it tends to reward inefficient shooting—the more a player shoots, the higher his value in EFF.

Player Impact Estimate (PIE) [12] is a method which evaluates a player's overall statistical contribution against the total statistics in the games he played in. Its formula is: (PTS + FGM + FTM - FGA - FTA + DREB + (.5 * OREB) + AST + STL + (.5 * BLK) - PF - TO) / (GmPTS + GmFGM + GmFTM - GmFGA - GmFTA + GmDREB + (.5 * GmOREB) + GmAST + GmSTL + (.5 * GmBLK) - GmPF - GmTO), in which a notation with prefix *Gm* means the game total of all players in that specific statistics matter. Although PIE weighs the player's contribution based on the team's overall statistics, similar to EFF, PIE doesn't reflect the player's role and tends to reward inefficient shooting.

Another measure adopted by ESPN, *Real Plus-Minus* (RPM) [13] focuses on showing a player's impact on the court. If his team is outscoring its opponent, the player contribution is "plus", otherwise, his contribution is "minus". RPM helps reveal the effectiveness of light scoring players, but it can also be biased since the statistics is heavily influenced by the player's on-court teammates.

Among all the metrics, *Player efficiency rating* (PER), developed by John Hollinger [14], is the most widely recognized one number indicator to measure the player' overall performance. PER not only takes into accounts of the players basic statistics such as points and assists, but also measures the player's per-minute and pace-adjusted performance. The formulas for PER calculation include [14]:

$$factor = \frac{2}{3} - \left(0.25 \times \frac{lgAST}{lgFGM}\right) \div \left(\frac{lgFG}{lgFTM}\right) \quad (1)$$

$$VOP = \frac{lgPTS}{lgFGA - lgOREB + lgTO + 0.44 \times lgFTA} \quad (2)$$

$$DREBP = \frac{lgREB - lgOREB}{lgREB} \quad (3)$$

$$\begin{aligned} uPER = \frac{1}{min} \times \Bigg(& 3P - \frac{PF \times lgFTM}{lgPF} + \\ & \left[\frac{FTM}{2} \times \left(2 - \frac{tmAST}{3 \times tmFGM}\right)\right] + \\ & \left[FGM \times \left(2 - \frac{factor \times tmAST}{tmFGM}\right)\right] + \frac{2 \times AST}{3} + \\ & VOP \times \Bigg[DREBP \times (2 \times OREB + BLK - \\ & 0.2464 \times [FTA - FTM] - [FGA - FGM] - \\ & REB) + \frac{0.44 \times lgFTA \times PF}{lgPF} - (TO + OREB) + \\ & STL + REB - 0.1936(FTA - FTM)\Bigg]\Bigg) \end{aligned} \quad (4)$$

$$PER = \left(uPER \times \frac{lgPace}{tmPace}\right) \times \frac{15}{lguPER} \quad (5)$$

In the above formulas, a notation with prefix *tm* means the team instead of player statistics with regard to that specific indicator, and a notation with prefix *lg* means the league statistics with regard to that specific indicator.

2.2. Statistics of Starters and Bench Players

Basketball is a team game. Anytime on court, a squad of five players playing at different positions including center (C), power forward (PF), small forward (SF), point guard (PG), and shoot guard (SG) collaborate closely to compete with the opponents both offensively and defensively. Typically, a NBA team consists of fifteen players, with five starters and ten bench players. The starters are the elite players of the team, among whom are the super stars like Michael Jordan, LeBron James, Stephen Curry, etc. The bench players are the second unit of the team whose roles are mainly to rotate the starters for breaks, or to substitute the starters if they are injured, or to enhance certain playing styles according to the coach's on-field strategies. Usually, during the 48 minutes of a game, the starters play between 25 to 40 minutes, and the bench players' minutes vary from a few to 20 minutes. Since the starters always compete with other teams' elite players, they make the most contributions to the team's overall performance. Accordingly, they are highly paid. The total salary of the five starters often accounts for 70 percent or even higher of the total team salary. As for the bench players, their contributions can also be further categorized: normally, five out of the ten bench players form the main rotation unit of a team. They play solidly in each game for ten to twenty minutes. The remaining bench players are composed of rookies, veterans,

and less talented players. These players' salary amount is small and sometimes negligible to the team salary total. Many times, they only play a few minutes per game. Hence, their statistics are usually very low and their contribution has little impact to the team's performance, In fact, in a typical NBA game, both teams will only rotate about ten players to compete and the rest players are in "did not play" (DNP) status. Due to these reasons and for the purpose of simplifying the simulation, in this work, we assume that *a team consists of only five starters and five main bench players* (i.e., one starter and one bench player for each of the five positions). The research problem is to *optimize the ten player selection with limited salary budget*.

2.3. Player Statistics and Team Performance

In NBA, a team's performance directly affects the franchise business running. A winning team, especially the national champion, can gain huge commercial advantages over those low performed teams. The players also benefit from the team performance in that they can get better salary offers and sign more endorsement contracts. A team's statistics is the summation of all its players' statistics. Intuitively, the higher the individual players' statistics, the higher the team's statistics and the winning rate, and the optimal player selection strategy would be to maximize the team winning percentage by combining the players' statistics.

Based on this assumption, we first examined the relationship between the team players' PER values and the team's winning rate. As mentioned in section 2.2, the starters of a team contribute the most to the team's overall performance. Hence, the *hypothesis* is that the total of the starters PER values will be proportional to the teams' winning rate and the two indicators should show strong correlation. However, data analytics showed the hypothesis is not accurate. For instance, Table 2 presents our analysis result of the 15 east conference NBA teams based on the 2016-17 regular season data collected from the official websites of NBA and ESPN [15], [16], [17]. The data is sorted according to the team starters' PER total. The team's season winning rate, the expected ranking according to PER, and the real ranking in the conference are shown in the last three columns. From the table, we can see that the PER value has positive correlation with the team winning rate and conference ranking, but the relationship is not significant. The Pearson coefficient of the PER total and the team's winning rate is 0.698.

The inaccuracy of the hypothesis might stem from not including the bench players. So we further examined the relationship between all ten players' PERs and the team's performance. Table 3 presents the result of the 15 east conference NBA teams. Again, the data is sorted based on the total of all ten players' PERs. Although the result still demonstrates positive correlation, the significance is even weaker (Pearson coefficient = 0.499).

Since the starters always compete with the elite players of the opponent team, and most time, the bench players compete with the second unit of the opponent team. Their statistics are actually measured in different standards and

TABLE 2. STARTERS' PER TOTAL AND TEAM PERFORMANCE

Team	PER	Win%	Expected	Real
Cleveland Cavaliers	94.8	0.622	1	2
Toronto Raptors	90.5	0.622	2	3
Washington Wizards	90.1	0.598	3	4
Boston Celtics	88.9	0.646	4	1
Chicago Bulls	86.9	0.500	5	8
Miami Heat	83.3	0.500	6	9
Indiana Pacers	83	0.512	7	7
Brooklyn Nets	82.2	0.244	8	15
Charlotte Hornets	82	0.439	9	11
Orlando Magic	81.4	0.354	10	13
New York Knicks	79.9	0.378	11	12
Milwaukee Bucks	79.8	0.512	12	6
Atlanta Hawks	78.5	0.524	13	5
Detroit Pistons	78.2	0.451	14	10
Philadelphia 76ers	76.6	0.341	15	14

TABLE 3. ALL TEN PLAYERS' PER TOTAL AND TEAM PERFORMANCE

Team	PER	Win%	Expected	Real
Milwaukee Bucks	158.0	0.512	1	6
Toronto Raptors	157.6	0.622	2	3
Charlotte Hornets	153.7	0.439	3	11
New York Knicks	152.7	0.378	4	12
Cleveland Cavaliers	150.8	0.622	5	2
Boston Celtics	146.8	0.646	6	1
Indiana Pacers	146.4	0.512	7	7
Washington Wizards	145.9	0.598	8	4
Chicago Bulls	143.4	0.500	9	8
Miami Heat	140.6	0.500	10	9
Detroit Pistons	139.2	0.451	11	10
Orlando Magic	138.0	0.524	12	13
Atlanta Hawks	137.7	0.354	13	5
Brooklyn Nets	137.6	0.244	14	15
Philadelphia 76ers	133.3	0.341	15	14

should be recorded separately, but the current NBA statistics system does not differentiate the different competition environments. All players' performance statistics are recorded using indicator values such as points, assists, etc. Bias is unavoidably introduced. A direct thought is to give different weights while evaluating individual player's contribution. In other words, adding different weights to the total calculations of the starters' PERs and bench players' PERs. The problem, however, cannot be easily solved by doing so. For example, Table 2 and Table 3 show that for Boston Celtics, both of the starters' PER total and all players PER total are not ranked top, but the team's winning rate is the highest in the conference. Another counterexample is Charlotte Hornets. Both of the starters' PER total and all players' PER total are ranked higher than the team's real ranking. This means simple regression cannot fix the problem. More variables need to be considered.

2.4. Offense Rating and Defense Rating

In a NBA game, the teams play on both offense side and defense side. Some teams play more offensively and some teams play more defensively. Any team with strength on only one side cannot guarantee winning. A general manager

TABLE 4. TEAM OFFENSE AND DEFENSE RATINGS VS. WINNING PERCENTAGE

Team	Off Rtg	Def Rtg	Diff	Win%
Golden State Warriors	113.2	101.1	12.1	0.817
San Antonio Spurs	108.8	100.9	7.9	0.744
Houston Rockets	111.8	106.4	5.4	0.671
Boston Celtics	108.6	105.5	3.1	0.646
Toronto Raptors	109.8	104.9	4.9	0.622
Utah Jazz	107.4	102.7	4.7	0.622
Los Angeles Clippers	110.3	105.8	4.5	0.622
Cleveland Cavaliers	110.9	108	2.9	0.622
Washington Wizards	108.5	106.9	1.6	0.598
Oklahoma City	105	105.1	-0.1	0.573
Memphis Grizzlies	104.7	104.5	0.2	0.524
Atlanta Hawks	102.3	103.1	-0.8	0.524
Milwaukee Bucks	106.9	106.4	0.5	0.512
Indiana Pacers	106.2	106.3	-0.1	0.512
Miami Heat	105.2	104.1	1.1	0.5
Chicago Bulls	104.6	104.5	0.1	0.5
Portland Trail Blazers	107.8	107.8	0	0.5
Denver Nuggets	110	110.5	-0.5	0.488
Detroit Pistons	103.3	105.3	-2	0.451
Charlotte Hornets	106.4	106.1	0.3	0.439
New Orleans Pelicans	103.3	104.9	-1.6	0.415
Dallas Mavericks	103.7	106.3	-2.6	0.402
Sacramento Kings	104.6	109.1	-4.5	0.39
Minnesota Timberwolves	108.1	109.1	-1	0.378
New York Knicks	104.7	108.8	-4.1	0.378
Orlando Magic	101.2	108.1	-6.9	0.354
Philadelphia 76ers	100.7	106.4	-5.7	0.341
Los Angeles Lakers	103.4	110.6	-7.2	0.317
Phoenix Suns	103.9	109.3	-5.4	0.293
Brooklyn Nets	101.9	108	-6.1	0.244

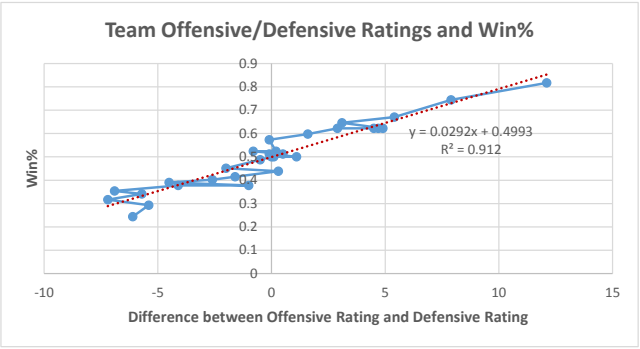

Figure 1. Team offensive and defensive rating difference versus winning percentage for regular season 2016-17

needs to consider balancing both sides while signing players. Nowadays, NBA uses two statistical indicators—offensive rating (Off Rtg) and defense rating (Def Rtg) to measure the team and player's respective performance in each side. Off Rtg is calculated by the points produced per 100 possessions by the team/player. Def Rtg is calculated by the opponent points allowed per 100 possessions. The higher the Off Rtg, the better the team or player's offensive skills. The lower the Def Rtg, the better the team or player's defensive skills.

Since different teams play in different styles, a true strong team can be measured by how many points they win on average over the opponents, (i.e., the difference between the team's Off Rtg and Def Rtg). Table 4 presents all the 30 NBA teams' statistics of the 2016-17 regular season [17]. After calculating the difference between the Off Rtg and Def Rtg, we can see that the result shows very strong correlation with the teams' winning rate. The Pearson coefficient between the two is 0.955. It also explained the reason why Boston Celtics ranked top in the east conference: they best outscore their opponents. Figure 1 confirms the proportional relationship between the two columns. The trendline depicts that the data can be well fitted with a linear regression model as follows:

$$Win\% = w_0 + (w_1 \times OffRtg) + (w_2 \times DefRtg) \quad (6)$$

3. Our Strategies and Optimization

3.1. Alternative Team Performance Formulation

Based on above analyses, the first objective of this research is to match individual players' efficiency with the team's statistics. We need effective player efficiency measures for player selection. Meanwhile, we require the measures precisely match the team's Off Rtg and Def Rtg to ensure team winning rate. Unfortunately, even though sports analysts and statisticians designed different player efficiency measures by combining basic NBA statistics, there is no such a single indicator satisfying both constraints. The main reason is the insufficiency of defense statistics in the current NBA statistics system. Compared to the relatively adequate offense statistics, more individual player's defense statistics should be recorded. The player Def Rtg, as discussed in Section 2.4, can be greatly impacted by the player's teammates. It cannot serve as a precise indicator by itself to estimate the player's efficiency.

In this research, we adopted two key indicators, player Def Rtg and PER, to match with the team's Def Rtg and the team's Off Rtg respectively. Player Def Rtg refers to the average of all players' Def Rtgs of a team. Since player Def Rtg is measured in a similar way to the team Def Rtg, our analysis shows that it can almost perfectly substitute the latter (Pearson coefficient = 0.978). In Section 2.3, we showed that PER cannot be used to accurately predict a team's performance, but our analysis showed it has strong correlation with the team's Off Rtg (Pearson coefficient = 0.83). In fact, the designer of PER, John Hollinger, freely admits that PER largely measures offense. Based on the two indicators' characteristics, we use a combination of PER and player Def Rtg as a tradeoff to comprehensively measure a player's performance.

To reformulate $eq.(6)$ using individual player's data, we made the following assumptions to simplify several basic NBA facts: (1) *a team consists of only five starters and five major bench players*; (2) *each player takes one fixed role on the court* (i.e., the player only plays at one fixed position out of point guard, shooting guard, small forward, power forward, and center); (3) *each player is pre-designated as either a starter or a bench player*; (4) *starters and bench players' contribution to the team should be weighted*; (5)

all players are free agents and their salary info is given. Correspondingly, a team's winning rate, as shown in $eq.(6)$, can be transformed to the following by using the team players' Def Rtgs and PERs:

$$\begin{aligned} Win\% = w_0 + (w_1 \times \sum tmStarterPER) + \\ (w_2 \times \sum tmBenchPER) + \\ (w_3 \times \sum tmStarterDefRtg) + \\ (w_4 \times \sum tmBenchDefRtg) \end{aligned} \tag{7}$$

If a logistic regression model is used, $eq.(6)$ will be transformed to:

$$\begin{aligned} y = w'_0 + (w'_1 \times \sum tmStarterPER) + \\ (w'_2 \times \sum tmBenchPER) + \\ (w'_3 \times \sum tmStarterDefRtg) + \\ (w'_4 \times \sum tmBenchDefRtg) \end{aligned} \tag{8}$$

$$Win\% = 1/(1 + e^{-y}) \tag{9}$$

It should be noted that if any of the assumptions needs to be changed for practical application purpose, the models can be easily adjusted to fit the new scenarios. For example, if a general manager wants to keep eight current team players and select two players to improve the team performance, he only needs to set eight PERs and eight Def Rtgs as constants by replacing them with those eight players' PER and Def Rtg values and leave the remaining as variables.

3.2. Strategies and Optimization

The goal of this research is to find which player can be a valuable addition to the team and how to sign the right ones to form a team within limited budget (e.g., salary cap). After the weight coefficients in the linear regression model $eq.(7)$ or the logistic regression $eq.(9)$ are fine-tuned, we can use them to formulate the optimization problem according to different player selection strategies. Based on the assumptions aforementioned, we first preprocess the collected NBA data. The players are categorized into ten individual groups based on the positions and the starter/bench status they typically play. Then the players are indexed from 1 to $10n$ where n is the number of players in each group (i.e., n practically equals the team number since we assume in each team there are only one starter and one bench player at each position). Specifically, Table 5 shows the player index, position, and starter (S)/bench (B) status information. Table 6 lists the variables used in formulating the strategies.

3.2.1. Strategy to Maximize Team Winning. This strategy is to select players to achieve the maximum team winning percentage under limited budget, which can be outlined as the objective function (10) and constraints (11)—(22). The objective function is derived from the team winning rate formula, $eq.(6)$. All notation values except S_i can be calculated or obtained directly from NBA and ESPN websites. This is an integer linear programming problem which is theoretically NP-hard. In our experiments, we leveraged the Python *PuLP* library for coding and computation. It should be noted that in the optimization model, since we predefined the player's starter/bench status, it means we assume in the new team after a player is selected, his contribution to the team's winning rate prediction is calculated based on his original position and starter/bench status. Extra variables can be introduced to enhance the model's flexibility, but the model will no longer be linear programming and we will not discuss it in this paper.

TABLE 5. PLAYER INDEX, POSITION, AND STARTER STATUS

Player Index	Position	S/B
1 to n	C	S
$n+1$ to $2n$	SF	S
$2n+1$ to $3n$	PF	S
$3n+1$ to $4n$	PG	S
$4n+1$ to $5n$	SG	S
$5n+1$ to $6n$	C	B
$6n+1$ to $7n$	SF	B
$7n+1$ to $8n$	PF	B
$8n+1$ to $9n$	PG	B
$9n+1$ to $10n$	SG	B

TABLE 6. VARIABLES AND CONSTANTS FOR OPTIMIZATION

Variables	Description	Type
$Salary_i$	Salary of the ith player	real
PER_i	the ith Player 's PER value	real
$DefRtg_i$	ith Player's Defense Rating	real
S_i	Selection status of the ith player	bool
w_i	Weight coefficients learned from regression	real
$SalaryCap$	Upper bound of team budget	real
$WinThreshold$	Minimum of team winning rate	real

$$\begin{aligned} \text{maximize} \quad & w_0 + w_1 \times \textstyle\sum_{i=1}^{5n}(S_i \times PER_i) + \\ & w_2 \times \textstyle\sum_{i=5n+1}^{10n}(S_i \times PER_i) + \\ & w_3 \times \textstyle\sum_{i=1}^{5n}(S_i \times DefRtg_i) + \\ & w_4 \times \textstyle\sum_{i=5n+1}^{10n}(S_i \times DefRtg_i) \end{aligned} \tag{10}$$

$$\text{s.t.} \quad \textstyle\sum_{i=1}^{10n}(Salary_i \times S_i) \leq SalaryCap \tag{11}$$

$$\textstyle\sum_{i=1}^{n} S_i = 1 \tag{12}$$

$$\textstyle\sum_{i=n+1}^{2n} S_i = 1 \tag{13}$$

$$\textstyle\sum_{i=2n+1}^{3n} S_i = 1 \tag{14}$$

$$\textstyle\sum_{i=3n+1}^{4n} S_i = 1 \tag{15}$$

$$\textstyle\sum_{i=4n+1}^{5n} S_i = 1 \tag{16}$$

$$\textstyle\sum_{i=5n+1}^{6n} S_i = 1 \tag{17}$$

$$\textstyle\sum_{i=6n+1}^{7n} S_i = 1 \tag{18}$$

$$\textstyle\sum_{i=7n+1}^{8n} S_i = 1 \tag{19}$$

$$\textstyle\sum_{i=8n+1}^{9n} S_i = 1 \tag{20}$$

$$\textstyle\sum_{i=9n+1}^{10n} S_i = 1 \tag{21}$$

$$S_i \in \{0, 1\} \tag{22}$$

3.2.2. Strategy to Minimize Team Salary. In this strategy, the objective is to minimize the team's salary total while

maintaining a certain team performance. The strategy is suitable for those teams whose top target is to save cost. A bottom winning rate must be predefined for formulating the constraints. In our experiments, we set $WinThreshold$ to be the winning rate of the 2016-17 national championship team, Golden State Warriors (81.7%). Very similar to the first strategy, the optimization model of this strategy is an integer linear programming problem. It includes all constraints from (12) to (22). The only difference is the objective function is replaced by a minimization of team salary, $eq.(23)$, and constraint (11) is replaced by another inequality constraint (24).

$$\text{minimize} \quad \sum_{i=1}^{10n}(Salary_i \times S_i) \tag{23}$$

$$\begin{aligned} w_0 + w_1 \times \sum_{i=1}^{5n}(S_i \times PER_i) + \\ w_2 \times \sum_{i=5n+1}^{10n}(S_i \times PER_i) + \\ w_3 \times \sum_{i=1}^{5n}(S_i \times DefRtg_i) + \\ w_4 \times \sum_{i=5n+1}^{10n}(S_i \times DefRtg_i) \geq WinThreshold \end{aligned} \tag{24}$$

4. Evaluation and Analysis

In this research, we used Python to program the system, the *scikit-learn* libraries to do linear and logistic regressions, and *PuLP* to solve the linear programming problem. Statistical data of the players and teams of the past six regular season's from 2011 to 2017 was gathered from NBA and ESPN websites. Players' salary information was collected from Hoopshype [18]. Non-statistical data such as the player's on-court position, the player's role as starter or bench player, and so on was collected by examining the players' profiles and the games they participated in (i.e., box-scores). All data is stored in a PostgreSQL database.

While studying the correlation between the team's winning percentage and the players' PERs and Def Rtgs, we used the 2016-17 regular season's data as the training set, and the other five seasons' data as the testing set. Both the linear regression model $eq.(7)$ and the logistic regression model $eq.(9)$ were trained, through which we found that the weight coefficients in the linear regression model are: $w_0 = 2.040343$, $w_1 = 0.011658$, $w_2 = 0.005990$, $w_3 = -0.002605$, and $w_4 = -0.002851$; The weight coefficients in the logistic regression model are: $w_0' = 6.94405$, $w_1' = 0.050758$, $w_2' = 0.025760$, $w_3' = -0.011108$, and $w_4' = -0.013042$. The R^2 value between the predicted winning percentage and the real winning rate is 0.831 under the linear regression model and 0.845 under the logistic regression model. Thus, given a team's ten player roster, we convert $eq.(7)$ to $eq.(25)$ to predict the team's winning percentage. Similarly, $eq.(9)$ can also be converted to the one with the trained weight coefficients. These weight coefficients were used not only in predicting a team's winning rate, but also in optimizing player selection.

$$\begin{aligned} 2.040343 + 0.011658 \times \sum tmStarterPER + \\ 0.00599 \times \sum tmBenchPER + \\ -0.002605 \times \sum tmStarterDefRtg + \\ -0.002851 \times \sum tmBenchDefRtg \end{aligned} \tag{25}$$

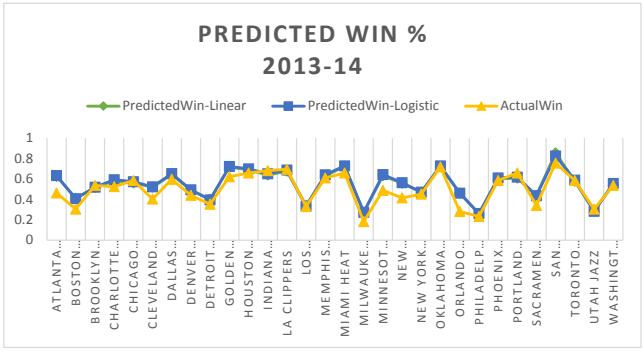

Figure 2. Predicted winning rate by linear and logistic regression models versus actual winning rate of regular season 2013-14

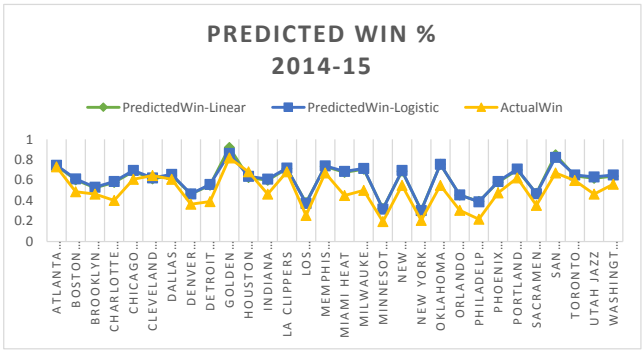

Figure 3. Predicted winning rate by linear and logistic regression models versus actual winning rate of regular season 2014-15

The precision of the regression models were validated using the other five regular seasons' data. Experimental results confirmed the prediction models were accurate for each season. Figures 2 to 4 depicted three seasons' results. We can see that the linear model and the logistic model have almost same prediction results. Both are very close to the real team winning percentage. The R^2 values for the linear model prediction versus the actual winning percentage are 0.840, 0.828, and 0.807 for 2013-14, 2014-15, and 2015-16 regular seasons respectively. The R^2 values for the logistic model prediction versus the actual winning percentage are 0.849, 0.833, and 0.814 for 2013-14, 2014-15, and 2015-16 regular seasons respectively.

After validating the accuracy of our regression models, we injected the trained weight coefficients into our player selection formulations. Both the "maximizing team winning percentage" and "minimizing team salary" strategies were tested for player selection. For the "maximizing team winning percentage" strategy, we set the $SalaryCap$ constraint to be the 2016-17 NBA salary cap \$94.143 million. For the "minimizing team salary" strategy, we set the $WinThreshold$ to be 81.7%—the regular season winning rate of the 2016-17 NBA championship team, Golden State Warriors. As discussed before, we assumed that all players from the 30 NBA teams are free agents and their current salaries will remain the same while signing them

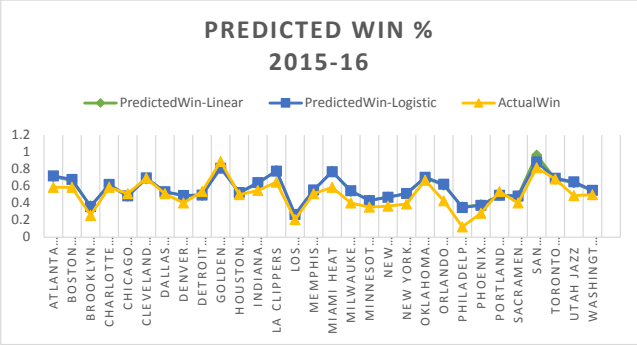

Figure 4. Predicted winning rate by linear and logistic regression models versus actual win rate of regular season 2015-16

to form a new team. These assumptions can be adjusted to fit practical use. For example, if a general manager is interested in signing some player, he can change the player's salary and replace it with the desired amount he wants to offer. Interested in seeing what kind of players a general manager can sign to form a team with the lowest winning rate under the same salary cap constraint, we also altered our "maximizing team winning" strategy by changing the objective from "maximizing" to "minimizing". The result can help reveal the worst contracts signed in the league and the players with the lowest performance-price ratios.

Table 7 presents the ten players selected to gain the "maximum team winning percentage" under the salary cap constraint. The results are same for both linear and logistic regression models. Requiring only a little NBA background knowledge, we can see that the players selection is effective and it ensures the team's talents on both the offense and defense sides. For example, Russell Westbrook, James Harden, and Kawhi Leonard ranked the top three in the 2016-17 season's most valuable player (MVP) voting. Joel Embiid and Kawhi Leonard are widely recognized as top defensive players. The team's total salary is $92,876,463.00. The logistic regression model predicts that the objective function value is 0.977, which means this team's winning percentage can be close to 98%. The linear regression model predicts that objective function is 1.362. This value means higher than 100% winning rate and is not practical. We can fix it by normalizing the results (skipped in this paper).

Table 8 presents the ten players selected to gain the "minimum team winning percentage" under the salary cap constraint. It should be noted that our formulation forces the selection of both starters and bench players since the goal is actually to find the worst contracts signed in the league. The results are same for both linear and logistic regression models. From the table, we can see that three players have only one-digit PER but signed with eight-digit or close salary. In fact, Timofey Mozgov, Chandler Parsons, and JR Smith were blamed frequently in the past season for their low performance-price ratios. The team's salary total is $93,706,394.00, even higher than that of Table 7. The logistic regression model predicts that the objective function value is 0.064, which means this team's winning percentage can be lower than 7%. The linear regression model predicts that objective function is -0.118399. Similarly, this value can be adjusted through normalization.

Table 9 presents the ten players selected to gain the "minimum team salary" under the team winning rate constraint. The results are same for both linear and logistic regression models. From the table, we can see that all players have solid PER and Def Rtg values, but none of them are paid high. Among these players, Rudy Gobert, Nikola Jokić, and Giannis Antetokounmpo are frequently referred as future stars. The reason they are low paid is because they are still within their rookie contracts. This team's salary total is only $13,087,629.00, less than one-seventh of the league's salary cap. The logistic regression model predicts that the objective function value is 0.860571, which means this team's winning percentage can be even higher than that of the national champion. The linear regression model predicts that objective function value is 0.830109.

5. Conclusion and Future Work

This research studied several important NBA player efficiency statistics and their impact on the team's winning rate. Two of the efficiency indicators, PER and Def Rtg, were combined to formulate the team's winning percentage prediction using linear regression and logistic regression models. Based on that, different player selection strategies were designed to achieve either "maximizing team winning rate" or "minimizing team salary". The optimization problems can be formulated as integer linear programming. The models and strategies were evaluated using the NBA data collected from the past six regular seasons. Experimental results showed that the models are accurate in predicting the team winning percentage and the strategies are effective in selecting the most appropriate players under different constraints. Future work includes exploring other models to gain higher prediction accuracy and designing new player selection strategies to better fit practical business running needs.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

[1] L. Steinberg, "Changing the game: the rise of sports analytics." [Online]. Available: https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/\#71e6207d4c1f

[2] J. Kubatko, D. Oliver, K. Pelton, and D. T. Rosenbaum, "A starting point for analyzing basketball statistics," *Journal of Quantitative Analysis in Sports*, vol. 3, no. 3, pp. 1–24, 2007.

[3] J. Piette, S. Anand, and L. Pham, "Evaluating basketball player performance via statistical network modeling," in *The 5th MIT Sloan Sports Analytics Conference*, 2011. [Online]. Available: http://www.sloansportsconference.com/?p=2840

TABLE 7. PLAYER SELECTION BASED ON PER AND DEFENSE RATING TO ACHIEVE HIGHEST TEAM WINNING PERCENTAGE

Name	S/B	Position	PER	Salary ($)	DefRtg	Current Team
Joel Embiid	S	C	24.2	4,826,160	99.1	Philadelphia 76ers
Nikola Jokić	S	PF	26.4	1,358,500	109.8	Denver Nuggets
Russell Westbrook	S	PG	30.7	26,540,100	104.6	Oklahoma City Thunder
Kawhi Leonard	S	SF	27.6	17,638,063	104	San Antonio Spurs
James Harden	S	SG	27.4	26,540,100	107.3	Houston Rockets
Jusuf Nurkić	B	C	21.1	1,921,320	103.7	Portland Trail Blazers
David Lee	B	PF	18.5	1,551,659	99.2	San Antonio Spurs
J.J. Barea	B	PG	17.3	4,096,950	102.8	Dallas Mavericks
Michael Beasley	B	SF	17.9	1,403,611	108.5	Milwaukee Bucks
Lou Williams	B	SG	24	7,000,000	109.2	Los Angeles Lakers

TABLE 8. PLAYER SELECTION BASED ON PER AND DEFENSE RATING TO ACHIEVE LOWEST TEAM WINNING PERCENTAGE

Name	S/B	Position	PER	Salary ($)	DefRtg	Current Team
Timofey Mozgov	S	C	12.3	16,000,000	109.0	Los Angeles Lakers
Domantas Sabonis	S	PF	6.9	2,440,200	101.9	Oklahoma City Thunder
Emmanuel Mudiay	S	PG	10.9	3,241,800	112.3	Denver Nuggets
Chandler Parsons	S	SF	7.7	22,116,750	107.6	Memphis Grizzlies
JR Smith	S	SG	8.1	12,800,000	108.1	Cleveland Cavaliers
Bismack Biyombo	B	C	12.2	17,000,000	108.2	Orlando Magic
Meyers Leonard	B	PF	9.0	9,213,484	107.6	Portland Trail Blazers
Cameron Payne	B	PG	6.2	2,112,480	103.7	Oklahoma City Thunder
Brandon Ingram	B	SF	8.6	5,281,680	112.6	Los Angeles Lakers
Brandon Rush	B	SG	6.6	3,500,000	109.8	Minnesota Timberwolves

TABLE 9. PLAYER SELECTION BASED ON PER AND DEFENSE RATING TO ACHIEVE MINIMUM TEAM SALARY

Name	S/B	Position	PER	Salary ($)	DefRtg	Current Team
Rudy Gobert	S	C	23.3	2,121,287	100.6	Utah Jazz
Nikola Jokić	S	PF	26.4	1,358,500	109.8	Denver Nuggets
Yogi Ferrell	S	PG	14.1	207,000	107.3	Dallas Mavericks
Giannis Antetokounmpo	S	SF	26.1	2,995,420	105.9	Milwaukee Bucks
Gary Harris	S	SG	16.5	1,655,880	114.0	Denver Nuggets
Salah Mejri	B	C	14.9	874,636	102.3	Dallas Mavericks
David Lee	B	PF	18.5	1,551,659	99.2	San Antonio Spurs
Spencer Dinwiddie	B	PG	12.7	45,000	106.3	Brooklyn Nets
Michael Beasley	B	SF	17.9	1,403,611	108.5	Milwaukee Bucks
Norman Powell	B	SG	14.1	874,636	105.4	Toronto Raptors

[4] J. Garcia, S. J. Ibanez, R. M. D. Santos, N. Leite, and J. Sampaio, "Identifying basketball performance indicators in regular season and playoff games," *Journal of Human Kinetics*, vol. 36, pp. 163–170, 2013.

[5] S. Sadiq and J. Zhao, "Money basketball: Optimizing basketball player selection using sas," in *Proc. of SAS Global Form*, no. 1790, 2014, pp. 1–9.

[6] S. Shea, *Basketball Analytics: Spatial Tracking*. CreateSpace Independent Publishing, 2014.

[7] B. Skinner and S. J. Guy, "A method for using player tracking data in basketball to learn player skills and predict team performance," *PloS One*, vol. 10, no. 9, 2015. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0136393

[8] L. Lamas, F. Santana, M. Heiner, C. Ugrinowitsch, and G. Fellingham, "Modeling the offensive-defensive interaction and resulting outcomes in basketball," *PloS One*, vol. 10, no. 12, 2015. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144435

[9] Y. Li and L. Ma, "Predict nba players career performance based on svm," in *Information Technology and Applications*. CRC Press, 2015, pp. 157–161.

[10] A. Franks, A. Miller, L. Bornn, and K. Goldsberry, "Counterpoints: Advanced defensive metrics for nba basketball," in *The 9th MIT Sloan Sports Analytics Conference*, 2015. [Online]. Available: http://www.sloansportsconference.com/content/counterpoints-advanced-defensive-metrics-for-nba-basketball/

[11] "NBA Salary Cap 2016-17." [Online]. Available: http://www.nba.com/2016/news/07/02/nba-salary-cap-set/

[12] "NBA Stats Glossary." [Online]. Available: http://stats.nba.com/help/glossary/

[13] "Real Plus-Minus." [Online]. Available: http://www.espn.com/nba/statistics/rpm/_/sort/RPM

[14] "Glossary." [Online]. Available: https://www.basketball-reference.com/about/glossary.html

[15] "NBA Player and Team Stats." [Online]. Available: http://stats.nba.com

[16] "NBA Player Statistics." [Online]. Available: http://insider.espn.com/nba/hollinger/statistics

[17] "NBA Team Stats." [Online]. Available: http://www.espn.com/nba/hollinger/teamstats

[18] "NBA Salaries." [Online]. Available: http://hoopshype.com/salaries/