**“Moneyball” in NBA to predict the performance of the Players**

Yu-Ching Kuo 5502

Krishna Chaitanya Sripada 5502

Kirill Novik 4502

**ABSTRACT**

The goal of this project was to apply moneyball to NBA teams based on the datasets obtained from [www.basketball-reference.com](http://www.basketball-reference.com). Moneyball is a term that refers to the practice or a strategy of assembling a maximally competitive sports team with minimal investments. In order to accomplish this goal, we analyzed the datasets and found the parameters that significantly contribute to the performance of a team. These parameters were picked according to the decision tree categorization that produced the least entropy, which then were combined into a performance index (Iperf). This Iperf was found to be able to describe if a team could make to playoff by an accuracy > 80%. Further, this index is applied to individual players and used to predict team performance for the 2014-15 season. With these attributes, we are able to correctly predict 5 teams (out of 8 teams) in the West conference and 6 teams (out of 8 teams) in the East conference of NBA league will be in the playoff. However, to achieve our ultimate goal which is to play moneyball in NBA, we need to look at both the performance and their salaries. By taking into account the salaries, we could find out those players who are underestimated. These are the players we need to form an economic but competitive NBA teams. We believe that the information provided here would be a good indication for team managers to make right decisions when the budget is limited.

**Keywords**

Basketball, Data Mining, Decision Trees, Moneyball, NBA, Performance, Performance Index, Prediction

**1. INTRODUCTION & METHODS**

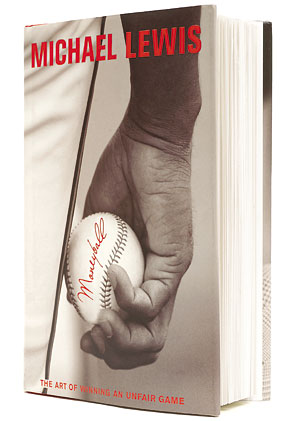


Figure 1,Moneyball: the Art of Winning an Unfair Game

Moneyball is a term that was first coined by Michael Lewis in his book “Moneyball: The Art of Winning an Unfair Game.” This book tells a story of the Oakland Athletics baseball team. The manager Billy Beane despite his disadvantaged revenue situation was able to assemble a competitive baseball team using an analytical, evidence-based, sabermetric approach.

According to the book, the wisdom collected in baseball over the past century is subjective and often flawed. Many statistical parameters such as stolen bases, runs batted, and batting average, typically used to evaluate players, are, in essence, artifacts of the 19th century that were never proven to directly contribute to the performance of a player or a team. The book argues that Oakland A’s managers were able to take advantage of statistical analysis to make a team that was able to compete against the richer competitors in Major League Baseball.

Lewis suggests that among the indicators of offensive success slugging percentage and on-base percentage are more precise and cheaper to obtain on the open market than speed and contact, the historically valued qualities. By following this strategy 2002 Athletics, with approximately $44 million in salary, were able to compete with such giants as the New York Yankees, with more than $125 million in salary. The 2002 Athletics were able to go to the playoffs for two consecutive years.[1]

Motivated by the story, we were questioning if moneyball can also be applied to the NBA teams. Not only would this approach help to maximize the performance  while minimizing payroll, but also would it help to estimate what teams are likelier to go the the playoffs. If used by team managers, this method could prevent them from providing big contracts to players who may not perform well in the following seasons.  These aspects motivated us to devise a statistical metric to guide the fielding of a highly competitive team with a minimal cost.

In order to be able to evaluate how well a player performs, the key factors contributing to the performance had to be determined and combined into a performance index (Iperf). The basketball reference website provides 24 metrics including:

Games (G), or how many games were played during the season, which may give some hints on how a particular player contributed to the team’s score for a certain season.

Games Started (GS), which is a metric that shows how many times during the season a player was on a starting lineup--an official list of players who play when the game begins, which is often comprised of the best players on the team. This metric may suggest that the player could be overestimated, or reveal certain performance patterns.

Minutes Played (MP), a straightforward metric that tells how many minutes were played by a player during the season.

Field Goals (FG), the term that refers to a per season basket scored on any shot or tap other than a free throw. [Wikipedia]

FGA (Field Goal Attempts), a metric that describes how many times a field goal was attempted throughout a season.

Field Goal Percentage (FG%),  which is a metric that refers to the ratio of field goals made to field goals attempted. Higher field goal percentage denotes higher efficiency. However, even very efficient players, such as guards, may have lower field goal percentage. We had to keep that nuance in mind to reduce bias when evaluating individual players.

3-Point Field Goals (3P), field goals made from beyond the three-point line, a designated arc surrounding the basket. A successful attempt is worth three points, in contrast to the two points awarder for shots made within the three-point line: a shot that requires great skill. This metric, taken in context of 3-Point Field Goal Attempts, may reveal how skillful a specific team is.

3-Point Field Goal Attempts (3PA), a metric that describes how many times a 3-Point Field Goal was attempted throughout a season. This metric, taken in context of 3-Point Field Goal, may reveal how skillful a specific team is.

3-Point Field Goal Percentage (3P%), which is a metric that refers to the ratio of 3-Point Field Goals made to 3-Point Field Goals attempted. Higher 3-Point Field Goal percentage denotes higher efficiency. However, usually players designated as point guards, shooting guards and small forwards are usually strong three-point shooters.  We had to keep this nuance in mind to reduce bias when evaluating individual players.

2-Point Field Goals (2P), field goals made from within the three-point line, a designated arc surrounding the basket. A successful attempt is worth two points.

2-Point Field Goal Attempts (2PA), a metric that describes how many times a 2-Point Field Goal was attempted throughout a season.

2-Point Field Goal Percentage (2P%), which is a metric that refers to the ratio of 2-Point Field Goals made to 2-Point Field Goals attempted. Higher 2-Point Field Goal percentage denotes higher efficiency. However, only certain positions are strong two-point shooters.  We had to keep this nuance in mind to reduce bias when evaluating individual players.

Effective Field Goal Percentage (eFG%), a metric used in NBA basketball that is similar to Field Goal Percentage.  This parameter adjusts for the fact that 3-point field goals are worth fifty percent more than 2-point field goals. eFG%=(FGM+(0.53PTM))/FGA. According to Sporting Charts, basic basketball statistics have long been unfair to players who shoot a lot of 3-pointers. Players who take and make a large number of 3-point shots often don't have a very impressive field goal percentage, but that doesn't tell the whole story in terms of the points that player produces.

Take, for example, a player like Hall of Fame center Bob Lanier. He generally did all his scoring around the basket. He had a career field goal percentage of 51.4%, but since he did all his scoring in the lane, his effective field goal percentage was also 51.4.% (http://www.sportingcharts.com/dictionary/nba/effective-field-goal-percentage-efg.aspx)

Free Throws (FT), or foul shots are unopposed attempts to score points from a restricted area on the court (the free throw line), and are generally awarded after a foul on the shooter by the opposing team. Each successful free throw is worth one point.

Free Throw Attempts (FTA), a metric that describes how many times a free throw was attempted throughout a season.

Free Throw Percentage (FT%), which is a metric that refers to the ratio of free throw made to free throw attempted. Only certain positions are strong free throwers.  We had to keep this nuance in mind to reduce bias when evaluating individual players.

Offensive Rebounds (ORB), metric that describes how many times a player was able to gain possession of the basketball after a missed field goal free throw. Offensive rebound denotes that the ball was recovered by the offensive side and did not change possession. This metric denotes how well a player can control the ball.

Defensive Rebounds (DRB), Offensive rebound denotes that the ball was recovered by the defensive side and did not change possession. This metric also denotes the ability of the player to control of the ball.

Total Rebounds (TRB), denotes the combination of offensive and defensive rebounds.

Assists (AST), a term denoting how many times a player was able to pass the ball to a teammate in a way that lead to a score by field goal.

Steals (STL), metric that refers to the ability of a defensive player to legaly cause a turnover by his positive, aggressive actions.

Blocks (BLK), occur when a defensive player legally deflects a field goal attempt from an offensive player.

Turnovers (TOV), occur when a team loses possession of the ball to the opposing team. This metric has a negative significance in the context of the performance of a player.

Personal Fouls (PF), term signifying breach of the rules that concern illegal personal contact with an opponent.

Points (PTS), metric that describes how many points a player scored during the season. [2][3]

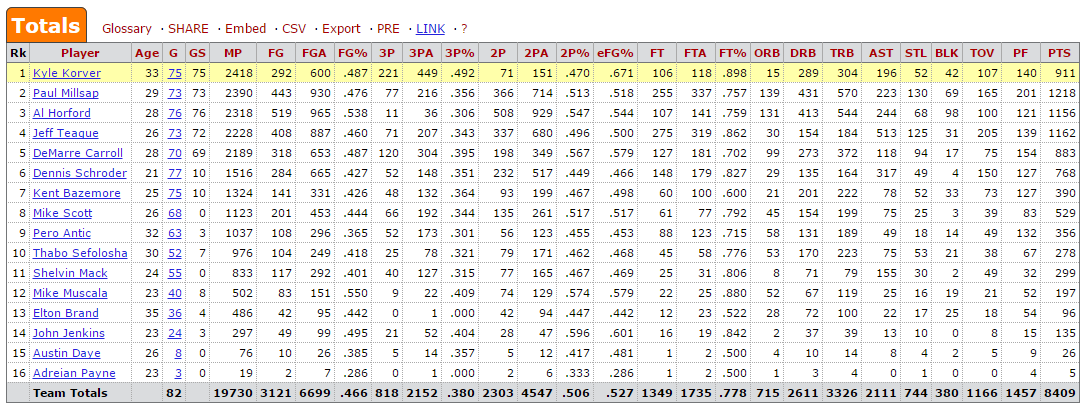


Figure 2 Team Totals Table from basketball-reference.com

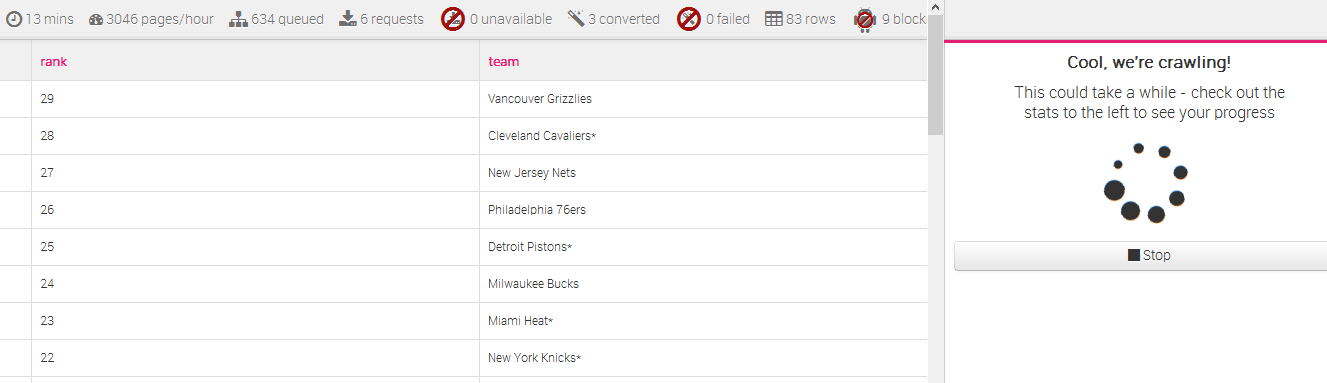
To be able to collect the attributes over a significant time frame, data could be mined either by using the web service queries of the website or, if this feature is absent a web-crawler could be utilized. We chose import.io as our web-crawler.  Import.io is a web-based UI platform for extracting data from websites. Users navigate to a website and teach the app to extract data by highlighting examples of data from the page, learning algorithms then generalize from these examples to work out how to get all the data on the website. The data that users collect is stored on import.io’s cloud servers and can be downloaded as CSV, Excel, Google Sheets or JSON and shared. Users can also generate an API from the data allowing them to easily integrate live web data into their own applications or third party analytics and visualization software. For more technical users, import.io offers real-time data retrieval through JSON REST-based and streaming APIs, integration with several common programming languages and data manipulation tools, as well as a federation platform which allows up to 100 data sources to be queried simultaneously.[4]

Figure 3 Import IO Crawler Output

Due to the large number of parameters, a data reduction step must be considered. To further determine the importance of attributes, we decided to construct a decision tree. A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Decision tree classification for this project uses information gain as its attribute selection measure. This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or “information content” of messages. Let node N represent or hold the tuples of partition D. The attribute with the highest information gain is chosen as the splitting attribute for node N. This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or “impurity” in these partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found. The expected information needed to classify a tuple in D is given by where pi is the probability that an arbitrary tuple in D belongs to class Ci and is estimated by . A log function to the base 2 is used, because the information is encoded in bits. Info(D) is just the average amount of information needed to identify the class label of a tuple in D. The information we have is based solely on the proportions of tuples of each class. Info(D) is also known as the entropy of D.

Now, suppose we were to partition the tuples in D on some attribute A having v distinct values. If A is discrete-valued, these values correspond directly to the v outcomes of a test on A. Attribute A can be used to split D into v partitions or subsets. These partitions would correspond to the branches grown from node N. Ideally, we would like this partitioning to produce an exact classification of the tuples. That is, we would like for each partition to be pure. However, it is quite likely that the partitions will be impure (e.g., where a partition may contain a collection of tuples from different classes rather than from a single class). How much more information would we still need (after the partitioning) in order to arrive at an exact classification? This amount is measured by . The term  acts as the weight of the jth partition. InfoA(D) is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected information (still) required, the greater the purity of the partitions.

Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is, . In other words, Gain(A) tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest information gain, Gain(A), is chosen as the splitting attribute at node N. This is equivalent to saying that we want to partition on the attribute A that would do the “best classification,” so that the amount of information still required to finish classifying the tuples is minimal (i.e., minimum InfoA(D)).[5]

After the appropriate classification is performed, in order to be able to merge these factors together into Iperf, some normalization must be applied. Because the min-max normalization preserves the relationships among the original data values, it was chosen as the data smoothing technique. Min-max zero-to-one normalization performs a linear transformation on the original data. Suppose that minA and maxA are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v, of A to v0 in the range [0;1] by computing [6]. Similar procedure could be performed for the individual players. First, players had to be categorized into several categories dependent on their positions such as power forward, small forward, center, shooting guard and point guard. These categories were used to adjust the weights for the parameters factored into the Iperf. These adjusted parameters could be divided by the salary of the player to determine if a player is overestimated or underestimated. We further could select players with top Iperf and combine them into a team trying to maximize Iperf.

**2. RESULTS & DISCUSSIONS**

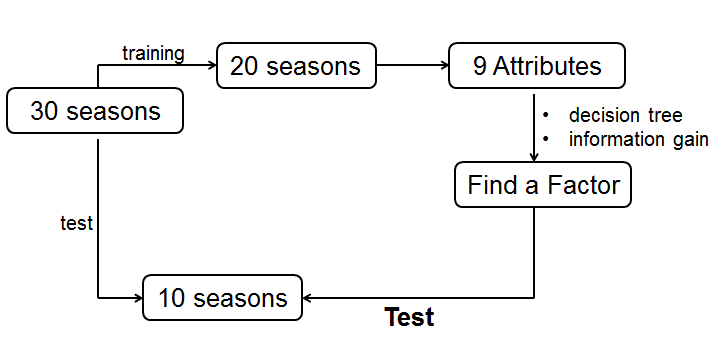
* 1. **Team Evaluation**

The first goal of this project is to assess the goodness of a team based on whether it goes to playoff or not. By finding the attributes that represent good teams we can further achieve our primary goal which is evaluating the performance of individual players. To accomplish this task, we need to identify parameters that significantly affect the performance. The datasets we are working with have 26 parameters to choose from, which after data reduction becomes only 9. A decision tree approach is applied to find out the most important attributes form the 9 attributes we choose. When the most important attributes are determined, we merge them into one single factor after applying the min-max normalization to these attributes and use it for describing the performance of a team.

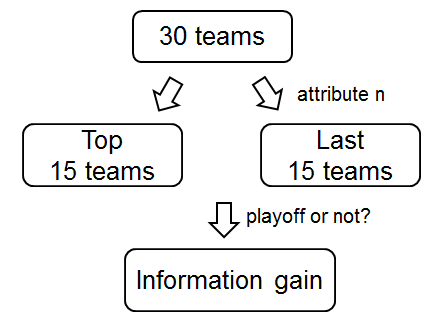
**Decision Tree**

From the 26 attributes, we reduce the attributes to 9 attributes including FG%, ORB, DRB, AST, STL, BLK, TOV, PF and Pts/G by ruling out the attributes that are either redundant or relevant. Here the decision is made based on if a team can make to playoff or not. Information gain for all the 9 attributes we choose are calculated to determine the importance of each attribute. We split the past 30 seasons of NBA league into 20 seasons as a training dataset and 10 seasons for model testing which has been shown in Scheme 1. To take into account for the temporal changes of the data, we take one season every other two seasons as our testing dataset while the rest of the seasons as the training dataset. In the work presented here, two approaches are used for determining the importance of the attributes.

1. Within the training dataset, we calculate the information gain for each attribute from the decision tree we construct in Scheme 2 based on the equations provided in the introduction for 20 seasons. By averaging the information gain overall these 20 years, the most critical attributes can be found. The results are shown in Table 1. As can be seen from Table 1, the top five most critical attributes that can determine the goodness of a team is FG%, DRB, TOV, Pts/G, and AST respectively due to their high information gain. The results indicate that to evaluate a player, points which is the most popular and easiest way to evaluate a player is not the best candidate. Instead, the field goal percentage and defense rebound are probably better candidates for player evaluation.



Scheme 1. Overview of the analytic model

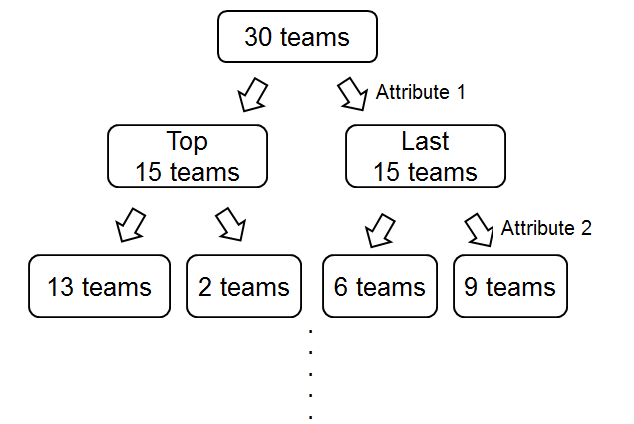


Scheme 2. 1 layer decision tree in this work

Table 1. The results of information gain in the past 20 seasons

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Information gain | Attributes | | | | | | | | |
| season | FG% | DRB | TOV | Pts/G | AST | ORB | BLK | PF | STL |
| 2013-2014 |  |  |  |  |  |  |  |  |  |
| 2012-2013 |  |  |  |  |  |  |  |  |  |
| 2010-2011 |  |  |  |  |  |  |  |  |  |
| 2009-2010 |  |  |  |  |  |  |  |  |  |
| 2007-2008 |  |  |  |  |  |  |  |  |  |
| 2006-2007 |  |  |  |  |  |  |  |  |  |
| 2004-2005 |  |  |  |  |  |  |  |  |  |
| 2003-2004 |  |  |  |  |  |  |  |  |  |
| 2001-2002 |  |  |  |  |  |  |  |  |  |
| 2000-2001 |  |  |  |  |  |  |  |  |  |
| 1998-1999 |  |  |  |  |  |  |  |  |  |
| 1997-1998 |  |  |  |  |  |  |  |  |  |
| 1995-1996 |  |  |  |  |  |  |  |  |  |
| 1994-1995 |  |  |  |  |  |  |  |  |  |
| 1992-1993 |  |  |  |  |  |  |  |  |  |
| 1991-1992 |  |  |  |  |  |  |  |  |  |
| 1989-1990 |  |  |  |  |  |  |  |  |  |
| 1998-1989 |  |  |  |  |  |  |  |  |  |
| 1986-1987 |  |  |  |  |  |  |  |  |  |
| 1985-1986 |  |  |  |  |  |  |  |  |  |
| average |  |  |  |  |  |  |  |  |  |

2. To further validate the five attributes found with approach 1, instead of one layer decision tree, we further split the decision tree into 5 levels and order the importance of the attributes as shown in Scheme 3. From the results presented in Table 2, the top five most critical attributes are found to be consistent with approach 1. They are FG%, DRB, TOV, Pts/G, and AST respectively. Therefore, the following analysis we will focus on these five attributes.



Scheme 3. multiple layers decision tree in this work

Table 2. The results of information gain for various layers of decision tree

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Information gain | Attributes |  |  |  |  |  |  |  |  |
| Layer | FG% | DRB | TOV | Pts/G | AST | ORB | BLK | PF | STL |
| 1st layer |  |  |  |  |  |  |  |  |  |
| 2nd layer |  |  |  |  |  |  |  |  |  |
| 3rd layer |  |  |  |  |  |  |  |  |  |
| 4th layer |  |  |  |  |  |  |  |  |  |
| 5th layer |  |  |  |  |  |  |  |  |  |

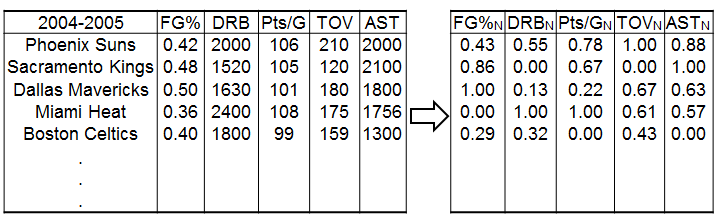
**Merge the attributes into a factor (Iperf)**

With the goal of using one single factor to describe a team, we need to merge several most critical attributes. Due to the variety of the scale for different attributes, a min-max normalization approach which is shown in eq 1 where the subscript N denotes normalized attribute, is applied to unify the scale of various attributes (scale are in 0~1 after the normalization) and make the merging applicable. An example of normalization is shown in Table 2 where we normalize the top five most important attributes in the season 2004-2005.

(1)



Table 2. An example of min-max normalization



After the normalization, we used eq 2 to merge the important attributes to generate a merged factor which we will name it Iperf to describe the performance of a team. However, it should be noted that the sign for the attributes TOV is opposite from the other attributes. This is because TOV contributes negatively to the performance of a team or a player which means the more TOV a team has, the worse performance should be expected. By eq 2, the new attribute (Iperf) can be found and it will be used to evaluate the NBA teams in each season. For each season, the 30 NBA teams will be split into the top 15 which have the highest value of Iperf and the last 15. In this work, we aim at finding an Iperf that can ensure 80% of the top 15 teams make to the playoff for the testing dataset.

(2)



To find the Attributen+1, the unknownparameters in eq 2 are required to be found. Since the scale of the normalized attributes are in 0~1, we postulate that the unknown parameters are ranging from 0 to 1. For each parameter, a trial and error approach with an interval of 0.01 from 0 to 1 is used for detecting the optimized parameter pair. Here the top three attributes are used for constructing the new factor due to its better accuracy compare to other models which can be found in Table 3. Within the 106 possible combinations of the unknown parameters, the optimized pair is found to be (1.0, 0.6, -0.2) with the training accuracy to be 81.3% for 20 seasons of the training datasets and the prediction accuracy to be 83.3% for the test datasets.

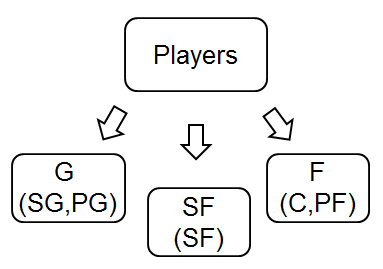
Table 3. The optimized parameters and accuracy for various models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number of parameters | a | b | c | Training  (20 seasons) | Testing  (10 seasons) |
| 1 parameter (a) | 1.0 | - | - | 0.736 | 0.773 |
| 2 parameters (a,b) | 1.0 | 0.6 | - | 0.79 | 0.827 |
| 3 parameters (a,b,c) | 1.0 | 0.6 | -0.2 | 0.813 | 0.833 |

* 1. **Individual Players**

**Issues and Challenges**

After defining the critical attributes and the important factor (Iperf) for evaluating teams, we aim at extending these findings to evaluate individual players. However, several issues should be noticed before applying the factor. First, the attributes such as DRB and TOV for a player are dependent on how many minutes he played for each. The more time a player stays on the court, the more likely for a player to grab rebounds and make turn overs. Normally, the average playing time varies a lot between players. Therefore, instead of looking at the attributes DRB and TOV, DRB/minutes and TOV/minutes are used instead for finding the key factor Iperf for players. Second, a basketball team consists of five different positions including center (C), power forward (PF), small forward (SF), shooting guard (SG) and point guard (PG). The FG% and DRB are generally higher and TOV is normally less for C and PF compare to other positions. Therefore, we need to categorize the players based on the position they played. Otherwise, the analysis would bias against SG and PG. To deal with the issue we split all the players into three categories: 1. Guards (G) including PG and SG, 2. Small forwards (SF) 3. Forwards (F) including C and PF. The categories are shown in Scheme 4. The last issue is to normalize the attributes before merging them into one single index. However, if we apply the normalization for all the players together, we would again bias the analysis. Therefore, we apply the min-max normalization for FG%, DRB/minutes and TOV/minutes in each category instead of whole players before calculating the index for players.



Scheme 4. Categories of the players

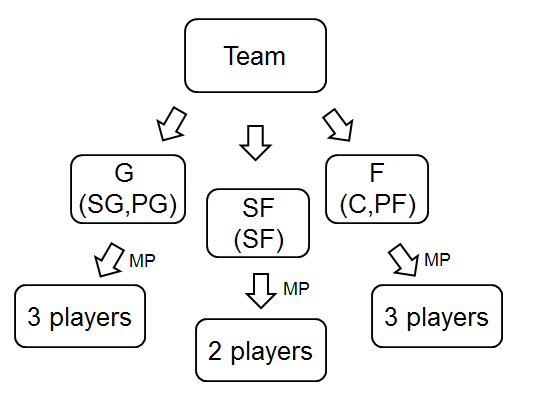
**Merging the Attributes for Individual Players**

Same approach like the last section is used for calculating the index for players. However, for players, how many minutes played is taken into account. Eq 3 where subscript N denotes normalized value is used instead of eq 2 used in the last section.

(3)



Via the same approach mentioned in the previous section, we optimized the parameters (d, e, f) for in eq 3 to find the key factor for players. To verify the accuracy of the description via the factor calculated here, we apply the factor to every player and sum up the factors for every team. By comparing the summation of the factors between teams, we would understand which teams are more likely to perform better and be in the playoff. Theoretically, the higher factor of a team, the better performance and higher probability to be in the playoff should be expected. However, there are two important issues that may bias the comparisons. First, the total number of players in each team varies significantly. The more players a team has, the higher score of the factors would be expected. Thus, the analysis is highly possible to introduce bias to big teams. Second, the number of players in various positions is different across the teams. For example, in the 2014~2015 season, Dallas Mavericks has two centers while Golden State Warriors has three centers. As mentioned above, normally C and PF have higher Iperf compared to the other positions. Therefore, the analysis would bias against those teams has less C and PF. To tackle with these issues, we don’t investigate all the players of teams. Instead, in each team, the top 3 guards, top 2 small forwards and top 3 power forwards who have the longer “Minutes Played per Game (MP)” would be taken into analysis. By doing so, the number of players and the number of players in each position are consistent across all the teams with total 8 players (3 guards, 2 small forwards and 3 power forwards (including centers)). The detailed scheme is shown in Scheme 5.



Scheme 5. The players involved in the analysis

Table 4 lists the summations of the Iperf for each team in the West conference and East conference of NBA league in the 2014-2015 season when the Iperf is obtained by the parameters(0.6, 0.3, -0.6). From the table, it can be seen that this model can predicts 5/8 of the teams in the West conference and 6/8 of the teams in the East conference make to the playoff (In NBA league, the top 8 players in each conference are qualified to be in the playoff.) This accuracy is 62.5% for the west conference and 75% for the east conference which are not perfect but acceptable in terms of its predictive nature. The predictions for individual players have much more issues which are discussed in the above sections compare to team evaluations. On the other hand, there are several important factors we may not be able to predict.

1. The health of players

Injury happened a lot in the NBA league and it is unpredictable. Healthy problems sometimes affect the performance of the team a lot especially when key players get injured. For example, Chris Bosh who is an all star power forward in Miami Heat got injured and missed the rest of this season which largely affects the performance of Miami Heat. Miami Heat ended up with 9th position in the East conference and didn’t make to the playoff.

1. Trade of players

NBA teams trade players in the middle of the season frequently. Sometimes, the deal may involve all star players and affect the ranking of the teams. The effect may be good or bad. For example, in 2014-2015 season, Dallas Mavericks sent several players to Boston and got all star point guard Rajon Rondo. However, this deal didn’t bring positive effect to Dallas. Dallas had only 31 wins but 24 loses after the trade which is worse than the 19 wins and 8 loses before the trade.

1. The regression of players

In the work presented here, we assume that the players would have similar performance as the previous season in order to simplify the model. To predict the regression of players is not a trivial work. There’s a literature [7] trying to predict the future performance of several players. However, the result is not good enough to estimate younger players like LeBron James. The regression of players is another possible factor to affect our prediction.

Table 4. The ranking of the NBA teams predicted by Iperf

|  |  |  |  |
| --- | --- | --- | --- |
| West Teams | Sum of Iperf | EestTeams | Sum of Iperf |
| SAS | 2.71 | MIA | 2.62 |
| LAC | 2.62 | DET | 2.58 |
| DAL | 2.47 | BRK | 2.48 |
| GSW | 2.43 | TOR | 2.31 |
| POR | 2.15 | WAS | 2.30 |
| PHO | 2.11 | CHO | 2.10 |
| MIN | 2.08 | ATL | 1.96 |
| DEN | 2.08 | CLE | 1.84 |
| MEM | 2.00 | ORL | 1.77 |
| NOP | 1.92 | IND | 1.72 |
| OKC | 1.86 | PHI | 1.59 |
| HOU | 1.83 | CHI | 1.52 |
| LAL | 1.83 | BOS | 1.46 |
| UTA | 1.56 | NYK | 1.34 |
| SAC | 1.41 | MIL | 0.91 |

* 1. **Performance versus Salaries**

With the goal of finding undervalued players, salaries must be taken into account. The idea of moneyball is to spend less amount of money but hire valuable players. Therefore, in the work here, we are not targeting those all star players such as Chris Paul, Stephon Curry and Kobe Bryant. We are targeting players who have high enough Iperf but require reasonable salaries. From Figure 4~8, the Iperf is plotted against the salaries. It can be noted that the players located in the right down corner are our candidates due to their relatively good performance and low salaries which denotes underestimated players. However, the players on the left up corner are the overestimated players that should be avoided.

Figure 4. Performance plot against player’s salaries (Centers)

Figure 5. Performance plot against player’s salaries (Power Forwards)

Figure 6. Performance plot against player’s salaries (Small Forwards)

Figure 7. Performance plot against player’s salaries (Point Guards)

Rasual Butler

Will Barton

Figure 8. Performance plot against player’s salaries (Shooting Guards)

Except for the figures, the detailed numbers of Iperf/salaries is provided in Table 5. which may be a good reference for evaluating NBA players. Depends on the budget and the strategy of a team, different decision may be made based on this table. Rich team may make decisions based on the value of Iperf while small team may put more focus on the value of Iperf/salaries to form a more economic and competitive team.

Table 5. The detailed performance and salaries of players

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | position | Iperf | USD(million) | Iperf/USD |
| AL\_HORFORD | C | 0.209 | 12.00 | 0.017 |
| AL\_JEFFERSON | C | 0.316 | 13.67 | 0.023 |
| ALEX\_LEN | C | -0.126 | 3.65 | -0.035 |
| AMARE\_STOUDEMIRE | C | 0.162 | 0.49 | 0.334 |
| ANDRE\_DRUMMOND | C | 0.566 | 2.57 | 0.220 |
| ANDREA\_BARGNANI | C | 0.016 | 11.50 | 0.001 |
| ANDREW\_BOGUT | C | 0.554 | 12.97 | 0.043 |
| ANTHONY\_DAVIS | C | 0.288 | 5.61 | 0.051 |
| BERNARD\_JAMES | C | 0.073 | 0.92 | 0.080 |
| BROOK\_LOPEZ | C | 0.219 | 15.72 | 0.014 |
| CHRIS\_ANDERSEN | C | 0.573 | 5.38 | 0.107 |
| CHRIS\_BOSH | C | 0.206 | 20.64 | 0.010 |
| CHRIS\_KAMAN | C | -0.071 | 4.80 | -0.015 |
| CODY\_ZELLER | C | -0.069 | 4.03 | -0.017 |
| DEANDRE\_JORDAN | C | 0.716 | 11.44 | 0.063 |
| DEMARCUS\_COUSINS | C | -0.112 | 13.70 | -0.008 |
| DWIGHT\_HOWARD | C | 0.166 | 21.44 | 0.008 |
| ENES\_KANTER | C | 0.043 | 5.69 | 0.008 |
| GORGUI\_DIENG | C | 0.16 | 1.41 | 0.113 |
| HASSAN\_WHITESIDE | C | 0.599 | 0.77 | 0.778 |
| HENRY\_SIMS | C | 0.128 | 0.92 | 0.140 |
| JOAKIM\_NOAH | C | 0.067 | 12.20 | 0.005 |
| JONAS\_VALANCIUNAS | C | 0.225 | 3.68 | 0.061 |
| JUSUF\_NURKIC | C | 0.025 | 1.76 | 0.014 |
| KELLY\_OLYNYK | C | -0.083 | 2.08 | -0.040 |
| KEVIN\_GARNETT | C | 0.105 | 12.00 | 0.009 |
| KOSTA\_KOUFOS | C | 0.214 | 3.00 | 0.071 |
| MARC\_GASOL | C | 0.08 | 15.83 | 0.005 |
| MARCIN\_GORTAT | C | 0.329 | 10.43 | 0.032 |
| MARREESE\_SPEIGHTS | C | -0.009 | 3.66 | -0.002 |
| NIKOLA\_PEKOVIC | C | 0.227 | 12.10 | 0.019 |
| NIKOLA\_VUCEVIC | C | 0.201 | 2.75 | 0.073 |
| OMER\_ASIK | C | 0.338 | 8.37 | 0.040 |
| PAU\_GASOL | C | 0.051 | 7.13 | 0.007 |
| ROBIN\_LOPEZ | C | 0.373 | 6.12 | 0.061 |
| ROY\_HIBBERT | C | -0.064 | 14.90 | -0.004 |
| RUDY\_GOBERT | C | 0.084 | 1.13 | 0.075 |
| SPENCER\_HAWES | C | 0.008 | 5.31 | 0.002 |
| STEVEN\_ADAMS | C | 0.084 | 2.18 | 0.038 |
| TIAGO\_SPLITTER | C | 0.176 | 9.25 | 0.019 |
| TIM\_DUNCAN | C | 0.128 | 10.00 | 0.013 |
| TIMOFEY\_MOZGOV | C | 0.13 | 4.65 | 0.028 |
| TYSON\_CHANDLER | C | 0.482 | 14.60 | 0.033 |
| ZAZA\_PACHULIA | C | -0.132 | 5.20 | -0.025 |
| AARON\_GORDON | PF | 0.07 | 3.99 | 0.018 |
| AMIR\_JOHNSON | PF | 0.229 | 7.00 | 0.033 |
| BLAKE\_GRIFFIN | PF | 0.069 | 17.67 | 0.004 |
| BORIS\_DIAW | PF | 0.086 | 7.50 | 0.011 |
| BRANDON\_BASS | PF | 0.15 | 6.90 | 0.022 |
| CARLOS\_BOOZER | PF | 0.026 | 3.25 | 0.008 |
| CARMELO\_ANTHONY | PF | -0.055 | 22.46 | -0.002 |
| CHANNING\_FRYE | PF | 0.092 | 8.18 | 0.011 |
| DAVID\_LEE | PF | 0.15 | 15.01 | 0.010 |
| DAVID\_WEST | PF | 0.12 | 12.00 | 0.010 |
| DERRICK\_FAVORS | PF | 0.191 | 12.83 | 0.015 |
| DIRK\_NOWITZKI | PF | 0.207 | 7.97 | 0.026 |
| DONATAS\_MOTIEJUNAS | PF | 0.058 | 1.48 | 0.039 |
| ED\_DAVIS | PF | 0.329 | 0.98 | 0.335 |
| ERSAN\_ILYASOVA | PF | 0.05 | 7.90 | 0.006 |
| GREG\_MONROE | PF | 0.115 | 5.48 | 0.021 |
| JARED\_SULLINGER | PF | -0.013 | 1.42 | -0.009 |
| JARNELL\_STOKES | PF | 0.186 | 0.73 | 0.257 |
| JASON\_THOMPSON | PF | 0.221 | 6.04 | 0.037 |
| JJ\_HICKSON | PF | 0.13 | 5.38 | 0.024 |
| JOHNNY\_O\_BRYANT | PF | -0.266 | 0.60 | -0.443 |
| JON\_LEUER | PF | 0.221 | 0.97 | 0.228 |
| JORDAN\_HILL | PF | 0.366 | 9.00 | 0.041 |
| KENNETH\_FARIED | PF | 0.226 | 2.25 | 0.100 |
| KEVIN\_LOVE | PF | 0.092 | 15.72 | 0.006 |
| KRIS\_HUMPHRIES | PF | 0.264 | 4.30 | 0.061 |
| LAMARCUS\_ALDRIDGE | PF | 0.197 | 16.01 | 0.012 |
| LEBRON\_JAMES | PF | -0.019 | 20.64 | -0.001 |
| LUIS\_SCOLA | PF | -0.002 | 4.87 | 0.000 |
| MARKIEFF\_MORRIS | PF | 0.011 | 2.99 | 0.004 |
| MARVIN\_WILLIAMS | PF | 0.161 | 7.00 | 0.023 |
| MASON\_PLUMLEE | PF | 0.406 | 1.36 | 0.299 |
| NENE\_HILARIO | PF | -0.029 | 13.00 | -0.002 |
| NERLENS\_NOEL | PF | 0.04 | 3.32 | 0.012 |
| OMRI\_CASSPI | PF | -0.027 | 1.06 | -0.025 |
| PATRICK\_PATTERSON | PF | 0.075 | 5.83 | 0.013 |
| PAUL\_MILLSAP | PF | -0.046 | 9.50 | -0.005 |
| PERO\_ANTIC | PF | -0.062 | 1.25 | -0.050 |
| PERRY\_JONES | PF | 0.2 | 1.13 | 0.177 |
| RYAN\_ANDERSON | PF | 0.132 | 8.49 | 0.016 |
| RYAN\_KELLY | PF | 0.077 | 1.65 | 0.047 |
| SERGE\_IBAKA | PF | 0.296 | 12.35 | 0.024 |
| TAJ\_GIBSON | PF | 0.021 | 8.00 | 0.003 |
| TERRENCE\_JONES | PF | 0.389 | 1.62 | 0.240 |
| THADDEUS\_YOUNG | PF | -0.073 | 9.41 | -0.008 |
| TOBIAS\_HARRIS | PF | 0.174 | 2.38 | 0.073 |
| TREVOR\_BOOKER | PF | 0.339 | 5.00 | 0.068 |
| TRISTAN\_THOMPSON | PF | 0.219 | 5.14 | 0.043 |
| ZACH\_RANDOLPH | PF | 0.026 | 16.50 | 0.002 |
| AVERY\_BRADLEY | PG | 0.407 | 7.19 | 0.057 |
| BRANDON\_JENNINGS | PG | 0.148 | 8.00 | 0.019 |
| BRANDON\_KNIGHT | PG | 0.27 | 3.55 | 0.076 |
| C\_J\_WATSON | PG | 0.343 | 2.11 | 0.163 |
| CHRIS\_PAUL | PG | 0.423 | 20.07 | 0.021 |
| DAMIAN\_LILLARD | PG | 0.311 | 3.34 | 0.093 |
| DANTE\_EXUM | PG | 0.147 | 3.62 | 0.041 |
| DARREN\_COLLISON | PG | 0.348 | 4.80 | 0.073 |
| DENNIS\_SCHRODER | PG | 0.155 | 1.69 | 0.092 |
| DERON\_WILLIAMS | PG | 0.323 | 19.75 | 0.016 |
| DERRICK\_ROSE | PG | 0.027 | 18.86 | 0.001 |
| DEVIN\_HARRIS | PG | 0.229 | 3.88 | 0.059 |
| ELFRID\_PAYTON | PG | 0.284 | 2.40 | 0.118 |
| ERIC\_BLEDSOE | PG | 0.358 | 12.17 | 0.029 |
| GEORGE\_HILL | PG | 0.457 | 8.00 | 0.057 |
| JAMEER\_NELSON | PG | 0.25 | 2.73 | 0.092 |
| JARRETT\_JACK | PG | 0.316 | 6.30 | 0.050 |
| JASON\_TERRY | PG | 0.209 | 5.85 | 0.036 |
| JEFF\_TEAGUE | PG | 0.215 | 8.00 | 0.027 |
| JOHN\_WALL | PG | 0.24 | 14.75 | 0.016 |
| JOSE\_CALDERON | PG | 0.403 | 7.10 | 0.057 |
| JRUE\_HOLIDAY | PG | 0.29 | 9.90 | 0.029 |
| KEMBA\_WALKER | PG | 0.305 | 3.27 | 0.093 |
| KIRK\_HINRICH | PG | 0.281 | 2.73 | 0.103 |
| KYLE\_LOWRY | PG | 0.329 | 12.00 | 0.027 |
| KYRIE\_IRVING | PG | 0.278 | 7.07 | 0.039 |
| LANGSTON\_GALLOWAY | PG | 0.424 | 0.24 | 1.798 |
| LOU\_WILLIAMS | PG | 0.273 | 5.45 | 0.050 |
| MARIO\_CHALMERS | PG | 0.322 | 4.00 | 0.081 |
| MICHAEL\_CARTER\_WILLIAMS | PG | 0.27 | 2.30 | 0.117 |
| MIKE\_CONLEY | PG | 0.336 | 8.89 | 0.038 |
| MO\_WILLIAMS | PG | 0.198 | 3.75 | 0.053 |
| MONTA\_ELLIS | PG | 0.28 | 8.36 | 0.033 |
| NORRIS\_COLE | PG | 0.279 | 2.04 | 0.137 |
| RAJON\_RONDO | PG | 0.279 | 12.91 | 0.022 |
| RAMON\_SESSIONS | PG | 0.368 | 2.08 | 0.177 |
| RAY\_MCCALLUM | PG | 0.278 | 0.82 | 0.340 |
| REGGIE\_JACKSON | PG | 0.381 | 2.20 | 0.173 |
| RICKY\_RUBIO | PG | 0.223 | 5.07 | 0.044 |
| RUSSELL\_WESTBROOK | PG | 0.258 | 15.72 | 0.016 |
| STEPHEN\_CURRY | PG | 0.287 | 10.63 | 0.027 |
| STEVE\_BLAKE | PG | 0.237 | 2.08 | 0.114 |
| TIM\_FRAZIER | PG | -0.013 | 0.05 | -0.256 |
| TONY\_PARKER | PG | 0.364 | 12.50 | 0.029 |
| TREY\_BURKE | PG | 0.244 | 2.55 | 0.096 |
| TY\_LAWSON | PG | 0.227 | 11.60 | 0.020 |
| VICTOR\_OLADIPO | PG | 0.239 | 4.98 | 0.048 |
| AL\_FAROUQ\_AMINU | SF | 0.572 | 0.98 | 0.583 |
| ANDRE\_IGUODALA | SF | 0.389 | 12.29 | 0.032 |
| BOJAN\_BOGDANOVIC | SF | 0.271 | 3.28 | 0.083 |
| CHANDLER\_PARSONS | SF | 0.376 | 14.70 | 0.026 |
| CHRIS\_COPELAND | SF | 0.336 | 3.14 | 0.107 |
| CLEANTHONY\_EARLY | SF | 0.1 | 0.51 | 0.197 |
| COREY\_BREWER | SF | 0.318 | 4.70 | 0.068 |
| DANILO\_GALLINARI | SF | 0.21 | 10.85 | 0.019 |
| DEMARRE\_CARROLL | SF | 0.511 | 2.44 | 0.209 |
| DERRICK\_WILLIAMS | SF | 0.421 | 6.33 | 0.066 |
| DORELL\_WRIGHT | SF | 0.355 | 3.15 | 0.113 |
| DRAYMOND\_GREEN | SF | 0.38 | 0.92 | 0.415 |
| EVAN\_TURNER | SF | 0.083 | 3.28 | 0.025 |
| GIANNIS\_ANTETOKOUNMPO | SF | 0.185 | 1.87 | 0.099 |
| GORDON\_HAYWARD | SF | 0.044 | 14.75 | 0.003 |
| HARRISON\_BARNES | SF | 0.294 | 3.05 | 0.096 |
| HENRY\_WALKER | SF | 0.256 | 0.12 | 2.218 |
| JABARI\_PARKER | SF | 0.294 | 4.93 | 0.060 |
| JAE\_CROWDER | SF | 0.453 | 0.92 | 0.495 |
| JEFF\_GREEN | SF | 0.184 | 9.45 | 0.019 |
| JEFFERY\_TAYLOR | SF | 0.138 | 0.92 | 0.151 |
| JERAMI\_GRANT | SF | 0.024 | 0.88 | 0.027 |
| JOE\_INGLES | SF | 0.141 | 0.51 | 0.278 |
| JORDAN\_HAMILTON | SF | 0.386 | 0.15 | 2.563 |
| JOSH\_SMITH | SF | 0.164 | 2.08 | 0.079 |
| JR\_SMITH | SF | 0.265 | 6.52 | 0.041 |
| KAWHI\_LEONARD | SF | 0.669 | 2.89 | 0.231 |
| KEVIN\_DURANT | SF | 0.241 | 19.00 | 0.013 |
| KHRIS\_MIDDLETON | SF | 0.265 | 0.92 | 0.290 |
| KOSTAS\_PAPANIKOLAOU | SF | -0.048 | 4.80 | -0.010 |
| KYLE\_ANDERSON | SF | 0.456 | 1.09 | 0.417 |
| KYLE\_SINGLER | SF | 0.35 | 1.09 | 0.321 |
| LANCE\_THOMAS | SF | -0.1 | 0.40 | -0.248 |
| LANDRY\_FIELDS | SF | 0.385 | 6.25 | 0.062 |
| LUC\_MBAH\_A\_MOUTE | SF | 0.55 | 4.38 | 0.125 |
| LUOL\_DENG | SF | 0.221 | 9.71 | 0.023 |
| MARCUS\_MORRIS | SF | 0.307 | 2.94 | 0.104 |
| MATT\_BARNES | SF | 0.363 | 3.40 | 0.107 |
| MAURICE\_HARKLESS | SF | 0.347 | 1.89 | 0.184 |
| MICHAEL\_KIDD\_GILCHRIST | SF | 0.501 | 5.02 | 0.100 |
| MIKE\_DUNLEAVY | SF | 0.347 | 3.33 | 0.104 |
| MIRZA\_TELETOVIC | SF | 0.406 | 3.37 | 0.121 |
| NICOLAS\_BATUM | SF | 0.317 | 11.77 | 0.027 |
| P\_J\_\_TUCKER | SF | 0.416 | 5.70 | 0.073 |
| PAUL\_PIERCE | SF | 0.243 | 5.31 | 0.046 |
| ROBBIE\_HUMMEL | SF | 0.515 | 0.88 | 0.585 |
| ROBERT\_COVINGTON | SF | 0.432 | 1.00 | 0.432 |
| RUDY\_GAY | SF | -0.026 | 19.32 | -0.001 |
| SHAWN\_MARION | SF | 0.537 | 1.45 | 0.371 |
| SOLOMON\_HILL | SF | 0.227 | 1.30 | 0.174 |
| T\_J\_WARREN | SF | 0.407 | 1.95 | 0.208 |
| TAYSHAUN\_PRINCE | SF | 0.452 | 7.71 | 0.059 |
| TREVOR\_ARIZA | SF | 0.401 | 8.58 | 0.047 |
| TYREKE\_EVANS | SF | 0.055 | 10.76 | 0.005 |
| WESLEY\_JOHNSON | SF | 0.378 | 0.98 | 0.385 |
| WILSON\_CHANDLER | SF | 0.334 | 6.76 | 0.049 |
| ALEC\_BURKS | SG | 0.366 | 3.03 | 0.121 |
| ANDRE\_DAWKINS | SG | -0.013 | - | - |
| ANDREW\_WIGGINS | SG | 0.351 | 5.51 | 0.064 |
| ANTHONY\_MORROW | SG | 0.454 | 3.20 | 0.142 |
| ARRON\_AFFLALO | SG | 0.41 | 7.75 | 0.053 |
| BEN\_MCLEMORE | SG | 0.302 | 3.03 | 0.100 |
| BRADLEY\_BEAL | SG | 0.355 | 4.51 | 0.079 |
| COURTNEY\_LEE | SG | 0.496 | 5.45 | 0.091 |
| DANNY\_GREEN | SG | 0.475 | 4.03 | 0.118 |
| DEMAR\_DEROZAN | SG | 0.373 | 9.50 | 0.039 |
| DION\_WAITERS | SG | 0.281 | 4.06 | 0.069 |
| DWYANE\_WADE | SG | 0.454 | 15.00 | 0.030 |
| ERIC\_GORDON | SG | 0.292 | 14.90 | 0.020 |
| EVAN\_FOURNIER | SG | 0.358 | 1.48 | 0.241 |
| GERALD\_GREEN | SG | 0.381 | 3.50 | 0.109 |
| GORAN\_DRAGIC | SG | 0.351 | 7.50 | 0.047 |
| IMAN\_SHUMPERT | SG | 0.389 | 2.62 | 0.149 |
| J\_J\_REDICK | SG | 0.405 | 6.58 | 0.062 |
| JABARI\_BROWN | SG | 0.265 | 0.04 | 5.920 |
| JAMAL\_CRAWFORD | SG | 0.245 | 5.45 | 0.045 |
| JAMES\_HARDEN | SG | 0.299 | 14.73 | 0.020 |
| JIMMY\_BUTLER | SG | 0.379 | 2.01 | 0.189 |
| JOE\_JOHNSON | SG | 0.43 | 23.18 | 0.019 |
| KENTAVIOUS\_CALDWELL\_POPE | SG | 0.406 | 2.77 | 0.146 |
| KEVIN\_MARTIN | SG | 0.362 | 6.79 | 0.053 |
| KLAY\_THOMPSON | SG | 0.381 | 3.08 | 0.124 |
| KOBE\_BRYANT | SG | -0.024 | 23.50 | -0.001 |
| KYLE\_KORVER | SG | 0.527 | 6.25 | 0.084 |
| LANCE\_STEPHENSON | SG | 0.551 | 9.00 | 0.061 |
| MANU\_GINOBILI | SG | 0.358 | 7.00 | 0.051 |
| MATTHEW\_DELLAVEDOVA | SG | 0.34 | 0.82 | 0.416 |
| O\_J\_MAYO | SG | 0.248 | 8.00 | 0.031 |
| PATRICK\_BEVERLEY | SG | 0.36 | 0.92 | 0.393 |
| QUINCY\_PONDEXTER | SG | 0.247 | 3.15 | 0.079 |
| RASUAL\_BUTLER | SG | 0.527 | 1.45 | 0.364 |
| RODNEY\_STUCKEY | SG | 0.302 | 1.23 | 0.246 |
| TERRENCE\_ROSS | SG | 0.421 | 2.79 | 0.151 |
| THABO\_SEFOLOSHA | SG | 0.452 | 4.15 | 0.109 |
| TONY\_ALLEN | SG | 0.441 | 4.83 | 0.091 |
| TONY\_SNELL | SG | 0.351 | 1.47 | 0.238 |
| TONY\_WROTEN | SG | 0.158 | 1.21 | 0.131 |
| VANDER\_BLUE | SG | 0.413 | - | - |
| WESLEY\_MATTHEWS | SG | 0.435 | 7.25 | 0.060 |
| WILL\_BARTON | SG | 0.513 | 0.92 | 0.561 |

1. **CONCLUSIONS**

To apply the concept of “Moneyball” to NBA, we need to figure out what attributes are the most important for describing the performance of a NBA team. Via decision tree approach and information gain calculations, the top five critical attributes are found to be FG%, DRB, TOV, Pts/G and AST. To achieve the goal of using a single index to predict the performance of a team, a min-max normalization is applied followed by merging the different attributes into one single index (Iperf). By trial and error, we determine the merging parameters to be (1, 0.6, -0.2). The accuracy of using top three attributes is >80% for both training datasets and test datasets. When comes to the individual player analysis, several issues including the attribute average varies across different positions and several attributes are time dependent. Therefore, the players are categorized into three groups (G, SF, PF) and normalized within the categories. With the same approach used for team analysis, the best parameters for individual player is (0.6, 0.3, -0.6). By the approach here, we are able to predict the ongoing 2014-2015 season by 2013-2014 player’s data. The predict accuracy are 5/8 for the West conference and 6/8 for the East conference. The error could come from multiple possibility including the health of the key players, trade in the middle of the seasons and the regression of key players. Finally, we take into account the salaries of the players and find out underestimated players. By doing the scatter plot and the comprehensive table provided in the discussion section, the underestimated and overestimated players are easy to found.

1. **References**

[1] Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. New York: W.W. Norton.

[2] Sporting Charts. (n.d.). Retrieved April 27, 2015, from <http://www.sportingcharts.com/dictionary/nba/effective-field-goal-percentage-efg.aspx>

[3] Basketball statistics. (n.d.). Retrieved April 27, 2015, from <http://en.wikipedia.org/wiki/Basketball_statistics>

[4] Instantly Turn Web Pages into Data. (n.d.). Retrieved April 27, 2015, from <https://import.io/>

[5] Han, J., & Kamber, M. (2006). Classification by Decision Tree Induction. In *Data mining concepts and techniques* (2nd ed., pp. 293-298). Amsterdam: Elsevier.

[6] Han, J., & Kamber, M. (2006). Classification by Decision Tree Induction. In *Data integration and transformation* (2nd ed., pp. 70-72). Amsterdam: Elsevier.

[7] Douglas Hwang. Forecasting NBA Player Performance using a Weibull-Gamma Statistical Timing Model, 2012.