SENG 474 - Spring 2015

Data Mining on NBA Statistics
Querying NBA Statistics for Match Outcomes and Player Counters.

Final Report

April 6th , 2015

Faculty Supervisor : Dr. Alex Thomo

Team : Erik Afable (V00692209)
Jonathan Lam (V00732001)
Sahibdeep Sran (V00486531)

Abstract

Data mining on sports has become very popular within the last decade. Analysis for predicting the outcome of a sporting event and selecting a competitive roster of players are 2 important issues in sports. With the prevalence of team and player statistics over the past decade, data mining enables us to approach these questions from a computational and statistical perspective. This paper presents one such statistical approach by querying statistical NBA team and player data to predict the outcome of a given match, predict how many points a player will score, determine the factors that influence if a particular player will score points on a shot, and select a defending player to counter a given offensive player. The problem of predicting the outcome of a given match and selecting a best defender to a given player are formalized as classification problems. Assessing the amount of points a player scores in a match is classified as a regression problem. These problems and their solutions are outlined in this report.

Keywords – NBA, basketball, sports.

I. Introduction

Sports analysis and prediction has garnered popularity among fans from around the world. This is particularly true for the more well known sports such as basketball. The National Basketball Association, or NBA, is the most prominent men's professional basketball league in North America. There are a myriad of stats for the teams and players comprising the NBA. Although many of the analysis requires subjective knowledge of the sport to make the best possible prediction of winners, this subjective knowledge can be influenced by emotion. By performing statistical analysis objectively on collected NBA data, we can perform predictions unbiased from these emotional influences. This paper presents several analyses performed on NBA statistical data and outlines their findings. Predicting the outcome of a given match and selecting a best defender are classification problems whereas predicting points per game from minutes played is answered using a Linear Regression.

In predicting a given match, the last 3 years of the Rocket data (comprising over 1000 tuples) was analysed and was processed by 7 classification models. In determining what influences the points scored by a player, we looked at the data of all Chicago Bulls players for the current season (comprising over 700 tuples) and performed a Linear Regression on various attributes to determine their influence on points. In selecting the best defender for an opposing player, we examined data pertaining to all of the shots that player LeBron James had taken in the current season (comprising over 1000 tuples)

and processed this data using the Naive Bayes classifier. The Linear Regression classifier was used to determine which attributes influence whether LeBron makes a shot.

We will now discuss the sources and process by which we obtained the data for our analysis.

II. Getting The Data

The data for this project were collected from two websites: Basketball-Reference.com and NBA.com. Basketball-reference contains all the basic statistics of each player, teams, seasons, and coaches. This information ranges from game statistics (W/L, Home/Away, etc.) to player salaries. Basketball-Reference.com provided us with most of the data that we used for our analysis. The data was easily accessible in .CSV format. The website also allowed queries of specific data sets which provided us with most of the data in a clean and usable format [1-4]. An example of a query is shown in Figure 2.1 below.

Figure 2.1: Basketball-Reference.com query [1].

The official NBA.com website on the other hand provided us with a set of more advanced statistics for each player such as time of possession, shot distance, and defender distance [5]. Unfortunately, the official NBA website only displayed these advanced statistics in forms of html tables on the website. NBA.com did not provide an easy way for us to extract the data. We manually copied the data from the NBA website and saved it into an Excel spreadsheet so we could analyze and clean it to be used with the Basketball-Reference.com data. Figure 2.2 on the next page shows a sample of the data available from NBA.com.

Figure 2.2: NBA.com advanced statistics data.

After gathering basic player and team data from Basketball-Reference.com and advanced player statistic data from NBA.com, we merged all of the information into a .CSV table for cleaning. This was easily performed through a simple Python script. An example of the output table is shown in Figure 2.3 below. We created an .ARFF to be processed in Weka for our data analysis.

Figure 2.3: An example tuple of NBA data we used in our analysis.

In the following sections we discuss the statistical analysis methods performed to answer our questions about the data.

III. Team Outcome Analysis

By looking at the data available from Basketball-Reference.com, we decided to do an analysis on predicting a certain team's game outcome against their opponent in the current season. With this in mind, we looked at the Houston Rockets team because they seemed to be a fairly stable team throughout the past couple years (i.e., there were no significant changes in their statistics data). Their top ranked player, James Harden, only missed one game in the current season and 8 games last season and 4 games the season before that. A top ranked player of a team is the player with the best statistics over a season such as games played, minutes played, field goal percentage, three point percentage, free throw percentages, assists, blocks, turnovers, and total points. James Harden had the highest scores compared with all of the other players on the team. The Houston Rockets seemed the most stable team for us to do our data-mining analysis on

team game outcomes. They were unlike other teams, where teams moved to another city or their top ranked player did not play for many games due to injuries.

We gathered the game statistics for each game of the last two seasons (2012-2013 and 2013 - 2014 season), because the 2012-2013 season was the season when James Harden got traded to the Rockets. We agreed that taking the data for the two seasons prior to the current season was enough. Many changes were happening for many NBA teams such as crucial trades of key players and team name changes (in some cases, a team changed their name to that of a previously existing team).

Figure 3.1: BayesNet Graph of Houston Rockets match outcomes.

The game results of each of the Houston Rocket games were predicted by looking at three attributes : Location of the game that it is played (home or away), the opponent that the Rockets are playing against, and if the top ranked player of the team played in the game or not. Figure 3.1 above shows the BayesNet graph produced by Weka. The graph shows the overall tree structure of how we are predicting the result of a game.

Classifier	**Results**
Naive Bayes	56.5%
J48	66.7%
ID3	58.0%
BayesNet	49.3%
Logistic	49.3%
Simple Logistic	58.0%

Table 1: Houston Rockets team outcome results.

We ran our data on a test set, which is the results of the games in the current 2014-2015 season. Table 1 above shows the results of several different classifiers that were used on our test set. The best results was J48, an algorithm that is used to generate a decision tree and it has a 66.7% of correctly predicting the game results.

IV. Points Analysis

Basketball is inherently a game of points, where the team with the most points wins a game. Therefore determining the factors that affect how many points a player on a team gets is vital. Thus, we wished to find out which factors influence how many points a player will get in the course of a game. To do this, we used the individual game stats for each player playing for the Chicago Bulls NBA franchise this current NBA season. The attributes for the game stats were as follows: minutes per game (MPG), 2 pointers attempted (2PA), 3 pointers attempted (3PA), free throws attempted (FTA), offensive rebounds (ORB), defensive rebounds (DRB), steals (STL), turnovers (TOV), personal fouls (PF), and points (PTS). We wanted to see which of the first 9 attributes (MPG to PF as mentioned above) were significant in determining the points a player gets. We used the Linear Regression classifiers in Weka and Microsoft Excel to determine our results.

When we used the Linear Regression Model in Weka, we obtained the results shown in Figure 4.1 below. There was a very high correlation (0.9218) between the equation obtained and our data points. We discovered that the attributes 2PA, 3PA, FTA, ORB, DRB, STL, and TOV all influenced the equation used to estimate a player's points.

Figure 4.1: Weka Linear Regression Classifier Output for Points Analysis.

When we used the Linear Regression Model in Microsoft Excel, we obtained the results shown in Figure 4.2 below. The P-values column helped us determine what factors (as

listed in the first column) influence a player's points. The lower an attribute's P-value was meant that attribute had a significant effect of the number of points a player scored in a game. Likewise, a higher P-value for an attribute indicates that attribute does not play a large factor in determining the points a player scores. Upon inspection, the 2PA, 3PA, and FTA by a player significantly affected how many points a player will score. The ORB, DRB, STL, and TOV attributes were minor factors in determining points. The MPG and PF were non factors in seeing how many points a player scored in a game.

Figure 4.2: Microsoft Excel Regression Statistics for Points Analysis.

Our group first believed that the minutes a player played per game would be significant in determining how many points a player got in a game. It seemed to be logical that the more minutes a player played, the more opportunities for points he would have. However, we didn't necessarily factor in that certain players may be better at taking shots and thus take more shots than other players. Even though both players may play the same amount of minutes.

Other than this, it seems quite obvious that the more shots a player attempts the more points he is likely to get. So it came as no surprise that 2PA, 3PA, and FTA had such a strong influence on the points per game a player got.

V. Specific Player Analysis

Ever since the inception of the NBA, statistical data (stats) has played a significant role in it. Stats have widely been used to measure which qualities a player possesses. Today, players are being tracked to the point where copious amounts of data about each individual player in the NBA is recorded. This makes data mining significant in order to decipher which stats are useful in determining the strengths and weaknesses of each player. Given the large amount of data that exists, we focused on exploring factors which may affect a player's shooting performance (whether a shot scores or misses) during a game. To do this, we selected one player in the NBA and used a data set detailing each shot he took during the course of the current NBA season. This data was then used to answer questions regarding whether a shot he takes hits (enters) or misses the basket. The questions we chose to ask using this data were as follows: which defender was closest to the player when he missed a shot and what attributes determine whether a player scores or misses a shot.

The NBA player we chose to analyze was LeBron James. The reasons for this were because LeBron is a 4 time NBA MVP, and is largely considered to be the best player in the NBA [6]. He is also one of the tops scorers in the NBA this season [7]. These facts naturally make it interesting to see what influences one of the league's best scorer's shots.

The first question we set out to answer was if LeBron James missed a shot, which opposing player (out of all the players he's played against this season) was closest to him. To do this we used the Naïve Bayes classifier on the shot result (whether a shot hits or misses the hoop) and closest player attributes of our data set. We created a new data set with only these two attributes and ran this data using the Naïve Bayes classifier in Weka. Weka returned that the player most likely to be closest to Lebron James when he misses a shot is Solomon Hill.

The next question was what attributes, in our data set, influenced whether LeBron James hit or missed a shot. We practically used the entire original data set (excluding the closest player attribute) in order to find this out. All the data we had was converted to be strictly numeric data and normalized; and we used the Linear Regression classifier in Microsoft Excel and Weka to find our conclusions.

The output of the Linear Regression in Weka is shown in Figure 5.1 on the next page. We determined that the attributes most relevant to him missing a shot were as follows: period, shot clock (the amount of time, in seconds, remaining on the shot clock when Lebron took the shot), touch time (the amount of time, in seconds, Lebron was in possession of the basketball), shot distance (the distance, in feet, Lebron was away from the hoop when he took the shot), and the defender distance (the distance, in feet, of the

closest defender when Lebron attempted his shot). However, the line formed from the attributes had a weak positive correlation with respect to the data set (0.3906). This meant that these attributes were not very influential in determining whether his shot hit or missed.

Figure 5.1: Weka Linear Regression Classifier Output for Player Analysis.

Upon examining the Linear Regression performed in Microsoft Excel (shown in Figure 5.2 on the next page), it was discovered that the P-values for most of the attributes were high, telling us there was no relationship between these values and whether LeBron hits or misses a shot. However the P-value for the shot distance attribute was significantly low allowing us to come to the conclusion that the distance LeBron is from the hoop can determine whether he hits or misses his shot. Upon further inspection of the data, the shot distance does seem to affect his shot: where his accuracy goes down the further he is from the hoop.

Figure 5.2: Microsoft Excel Regression Statistics for Player Analysis.

VI. Code Tools

Data was queried from Basketball-Reference.com and downloaded in a .CSV format. Python was used to perform simple pre-processing tasks such as clearing columns and formatting the data to be used in Weka. The .CSV format allowed us to easily convert the data into a form that is desired for querying in Weka (.ARFF). In Microsoft Excel, we used a built in data analysis feature (Regression) to analyze the perform a regression analysis. In Weka, we queried the data asking questions and performing our statistical analysis through several models.

VII. Conclusion

In this paper, we presented several analysis performed on NBA statistical data and outlined our findings. Predicting the outcome of a given match and selecting a best defender were classification problems, whereas predicting points per game from minutes played and determining which characteristics attributed to a successful shot were answered using a Linear Regression. We analysed over 1000 tuples each, for

predicting the outcome of a match between two teams, selecting the best defender, and predicting the points per game based on minutes played. Several classification models were used in the team prediction and found that the Bayes Network provided the lowest accuracy at 49.3% and J48 Decision Tree provided the highest accuracy at 66.7%. Naive Bayes was used to classify the best defender to LeBron James as Solomon Hill and we discovered that the distance from the net had the greatest impact on whether LeBron James scored a point. A Linear Regression was used to analyse what influenced a player's points per game and we found that the 2 pointers, 3 pointers, free throws were most significant but offensive and defensive rebounds, steals, and turnovers were also factors.

VIII. Discussion

We began our project seeking to ask questions to a set of NBA data and to perform statistical analysis using data mining techniques and answer these questions objectively. We used our team prediction model over two games that had recently been played and found our result to be correct. We also garnered interesting information in countering a specific players such as having identified Solomon Hill as the best defender on LeBron James. This data mining project provided a valuable learning experience to our group and allowed us to experiment with a data set that greatly interests us. For future work, we could take more attributes into account within our questions. Adding possible key attributes may help us get better and more accurate predictions. For example, when we predict the team outcome, we could add two possible key attributes: "travelled distance" and "is back to back game." The distance that a team needs to travel for their away game and if the team played a couple of games in a row may be a factor that can affect the game outcome because the players may be tired physically and mentally.

IX. References

[1] Basketball-Reference.com. (2015, Apr.). *Chicago Bulls Minutes Played and Points Query – Player Game Finder*. [Online] Available: http://www.basketball-reference.com/play-index/pgl_finder.cgi?request=1&player_id=&match=game&year_min=2015&year_max=2015&age_min=0&age_max=99&team_id=CHI&opp_id=&is_playoffs=N&round_id=&game_num_type=&game_num_min=&game_num_max=&game_month=&game_location=&game_result=&is_starter=&is_active=Y&is_hof=&pos_is_g=Y&pos_is_gf=Y&pos_is_f=Y&pos_is_fg=Y&pos_is_fc=Y&pos_is_c=Y&pos_is_cf=Y&c1stat=mp&c1comp=gt&c1val=0&c2stat=pts&c2comp=gt&c2val=0&c3stat=&c3comp=gt&c3val=&c4stat=&c4comp=gt&c4val=&is_dbl_dbl=&is_trp_dbl=&order_by=mp&order_by_asc=&offset=0

[2] Basketball-Reference.com. (2015). *Pau Gasol 2014-15 Game Log*. [Online] Available:

http://www.basketball-reference.com/players/g/gasolpa01/gamelog/2015/

[3] Basketball-Reference.com. (2015). *2012-13 Houston Rockets Schedule and Results*. [Online] Available: http://www.basketball-reference.com/teams/HOU/2013_games.html

[4] Basketball-Reference.com. (2015). *James Harden 2012-13 Game Log*. [Online] Available: http://www.basketball-reference.com/players/h/hardeja01/gamelog/2013/

[5] NBA.com. (2015). *LEBRON JAMES – 2014-15 REGULAR SEASON SHOOTING STATS*. [Online] Available: http://stats.NBA.com/player/#!/2544/stats/shooting/

[6] Ranker.com. (2015). *The Top NBA Players Of All Time*. [Online] Available: http://www.ranker.com/crowdranked-list/the-top-NBA-players-of-all-time?var=3&utm_expid=16418821-93.i9oekwBAQpeI307JNFGPBA.2&utm_referrer=http%3A%2F%2Fwww.google.ca%2Furl%3Fsa%3Dt%26rct%3Dj%26q%3D%26esrc%3Ds%26source%3Dweb%26cd%3D3%26ved%3D0CCkQFjAC%26url%3Dhttp%253A%252F%252Fwww.ranker.com%252Fcrowdranked-list%252Fthe-top-NBA-players-of-all-time%26ei%3DGO0iVb7zGJfXoATw74DQCw%26usg%3DAFQjCNGNnHeG6DTaP0u_1eDIblUbQ2utTA%26sig2%3DTYjvCa6bgFA-_CdUmo-Qzw%26bvm%3Dbv.89947451%2Cd.cGU

[7] NBA.com. (2015). *Official League Leaders -- 2014-15 REGULAR SEASON LEADERS*. [Online] Available: http://stats.NBA.com/leaders/#!?StatCategory=PTS