**Analysis of Players’ Salary by Performance Statistics**

**And**

**Prediction of NBA Final Champion by DEA Model**

Chen Zou, Zhiyuan Liu, Jiaqi Liu, Wen Liu

**Abstract**

This report describes the experiments conducted in case study 2 at DS 501 by group 8. The basic item of these experiments is collecting data from NBA Stats API. This report includes two major data collection: 1) using py-Goldsberry tool to get a list of NBA players’ information including names, IDs and numeric performance statistics. 2) shots moment data. With data collected, further research is conducted trying to emulate a player’s performance impact estimation (PIE) score and to predict his salary with it. The other research is to predict NBA final champion in 2014-2015 season by Data Envelopment Analysis (DEA).

**Key words**: NBA, Player Salary, Champion, DEA, PIE

**Motivation and Background Research**

Two topics are often referred to when people (no matter they are basketball lover, NBA player and team manager) talk about NBA: 1. who will win national championship this session? 2. Are player payed fairly?

Also interested in such topics, we decided to pick out the most competitive team in the league and try to compare a player’s performance with his annual salary. The first one gives us a chance to predict who will be the champion and the second one shows us which player is seriously undervalued.

This is a good time point for us to do prediction because the trade window has closed, players list won’t change.

In problem 4, we conducted experiments in data mining manner. First, we found the relationship between player’s overall performance and his salary estimation. Also, based on team statistical data in 2014-2015 regular season, we implemented an efficient model for measuring productive efficiency of decision making units (or DMUs) to evaluate the performance of each NBA team and estimated the performance score from the data.

**Realization**

**Problem 1: Collecting Data about NBA Players**

We use py-Goldsberry package to get general NBA player's information such as names and player's IDs, and then load the information into a data frame using pandas package. Py-Goldsberry is a really good tool to get data from stats.nba.com, while there is no public official API. We use PlayerList function to get player list of 2015 and load in dataframe. Finally we print the list to see the result.

**Problem 2:** **Shot Chart of the Player**

1. Select a player and print his ID

We choose Josh Smith as our player. We match his name in the dataframe and show his ID. Be aware that in stats.nba.com, the player name is shown as *Last Name, First Name. So we match Smith, Josh and get his id.*

1. Get the photo of the player from NBA stats website based upon the player id

Once we know Josh Smith's NBA ID, we can use it to find the url of his photo in stats.nba.com and scrape his picture using urllib library.

1. Collect the shot data of the player and print top 20 rows

We use nbastats library, which can get player's shot data very dicrectly. We write ShotChart function in nabstats, conveying Josh Smith' id as parameter, so now we have Josh's shot information. And we print top 20 rows of the list to see the details.

4) plot the player's shot data

Matplotlib and seaborn are two great packages to plot the data. As we can see from the shot information in the 3) list, there are shots' location features LOC\_X and LOC\_Y. So we can use these features to plot shot points in a court.

First we can draw these plots to see what they are look like. We notice that the x- axis values are the inverse of what they actually should be. So we will deal with that later. Then we begin to draw the court first. We will be using Matplotlib Patches's Circle, Rectangle and Arc objects. Importantly, we should plot the equally proportional court of real court. After we finishing the court we can draw the plot on it. Furthermore, we use jointplot fuction in seaborn to plot one kind of heat map of the shot. Which we can have a clear look of the shot distrbution of Josh Smith.

**Problem 3: Game Moments**

1. Show the animation of the moment of one game of this player

finding the 2014-15 season playoffs Houston VS. Clippers game 6 moments from stats.nba.com, we use the url to animate the moments. In this moment Josh Smith Grabbed the defence rebound and led a quick attack.

1. Download the movement data of all the players in this game event

We use requests package to download the data in this moment. And then we create a list containing the movement data for each player and the ball combining to data sets.

1. Plot the player's move

We can see how Josh was moving in that moment.

1. Compute basic statistics of the player.

We analyze the player's running distances and speeds in this moment, and also print the distances between players. We use euclidean function to compute the distance and using groupby method to apply to every player.

5) Close look at the score second

We compute draw the distance between ball and Josh to see when he was passing the ball and also see the distances between players during that period.

**Problem 4: Additional Experiments**

Experiment 1: Predict Players’ Salary Based on Their Performance Impact Estimations

By py-Goldsberry tool, we could get a list of NBA player's performance data. Is there strong a relationship between player's performance with their salary level? Hence, we want to conduct the experiment to find the relationship between player performance data and their salaries.

1. Get NBA career stats data for each season

We use Goldsberry API first get every player information about their name, NBA ID number, start time of getting NBA, and the contract end time. Then use their ID as index to find respective data of career stats in regular season. We get 23 items data set for each season. Because the duty of each players on the court is different, and the performance of them are various, if we only use the number of each player got in each season is biased. Hence to better describe the performance of each player, we employ PIE (player impact estimate) as key data.

PIE = (PTS + REB + AST + STL + BLK - (FGA - FGM) - (FTA-FTM) - PF)/MIN

PIE measures a player's overall statistical contribution against the total statistics in games they play in.

We also scrap each player salary in each regular season from 2008 to 2013 on <http://espn.go.com/nba/salaries>. Then combine the three parts of data (players’ information, PIE for each season and corresponding salary for that season) into a tidy data set for further analysis. The total number of observations is 850.

1. Analyzing the relationship between salaries and PIE

The distribution of PIE versus salary is hard to fit by any regression model. Hence, we estimated the distribution of mean salary under each PIE score. We found that the PIE and salary estimate had a linear relationship. We conducted a linear regression model between them and got a good fit result. We also performed cross validation to test this model. The figures are in the Result part.

Experiment 2: Predict NBA Final Champion in 2014-2015 season by DEA Model

Another interesting problem that people care most is that who will be the champion of each season. To solve this problem in data science manner, we introduced a performance evaluation criteria to predict the NBA final champion based on each team's performance data in regular season of 2014-2015. The reason of choosing 2014-2015 season is that NBA just started a few matches of pre-season's competition in this season, there isn't enough data for analysis. The other reason is that the competition results of 2014-2015 season have already came out, therefore we can compare the performance evaluation to the real result. The data analysis procedure is listed below.

1. Collecting team's performance data from NBA stat website

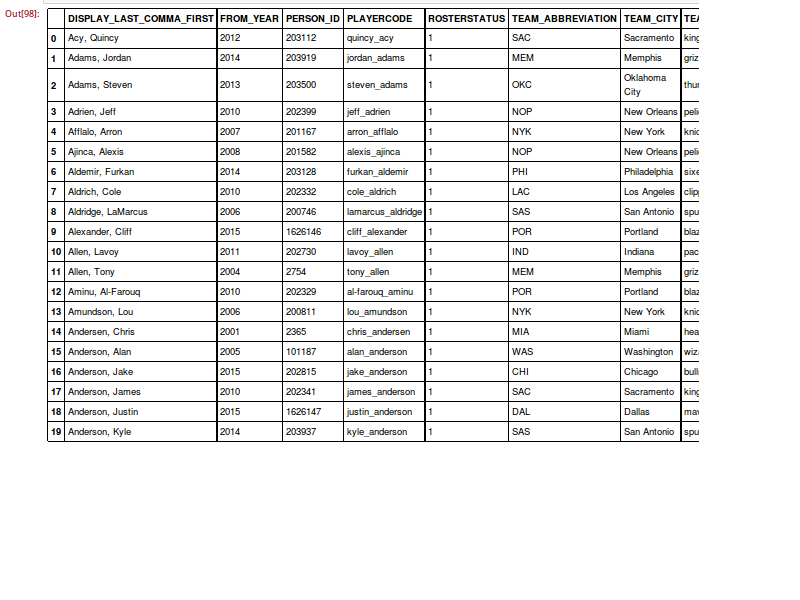
We collected 30 teams from east and west unions and their performance attributes of FGA, FG%, 3PA, 3P%, FTA, FT%, AST, STL, BLK, PTS, TOV and BLKA. The data is stored in a excel form.

1. Calculating each team's performance score by Data Evenlopment Analysis model in excel

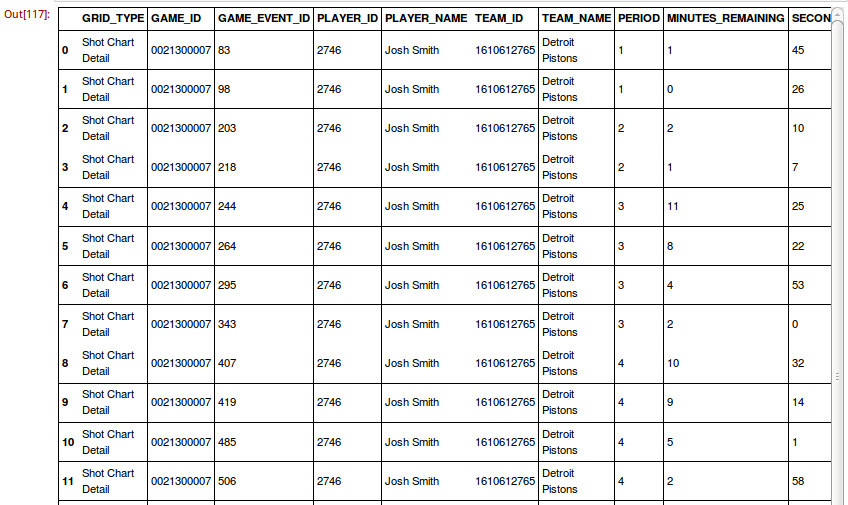
Data envelopment analysis (DEA) is a linear programming methodology to measure the efficiency of multiple decision-making units (DMUs) when the production process presents a structure of multiple inputs and outputs. For this problem, the measurements are the variables of team statistics and our decision is to choose the best team. We think the DEA model has the benefit of evaluating team's practical efficiency because it is designed for evaluating productivity and making desicions in operations research. We use the linear programming solver in excel to calulate the evaluation score of each team.

**Results**

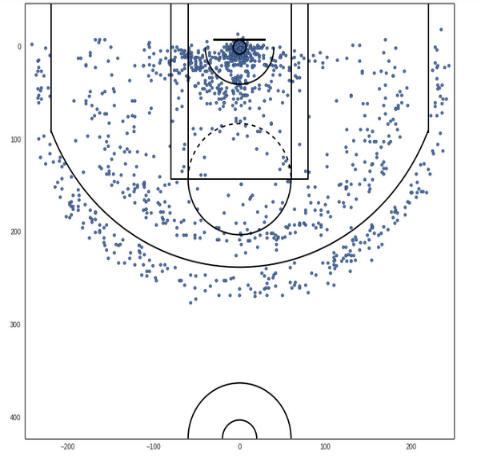
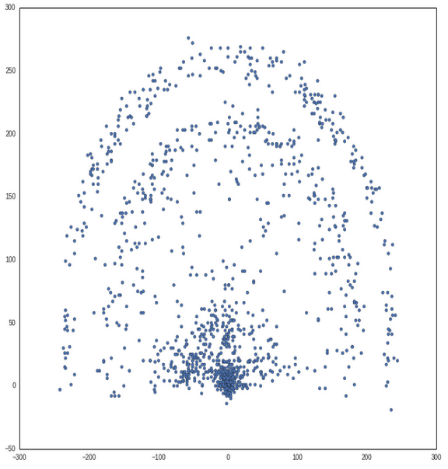
**Problem 1:**

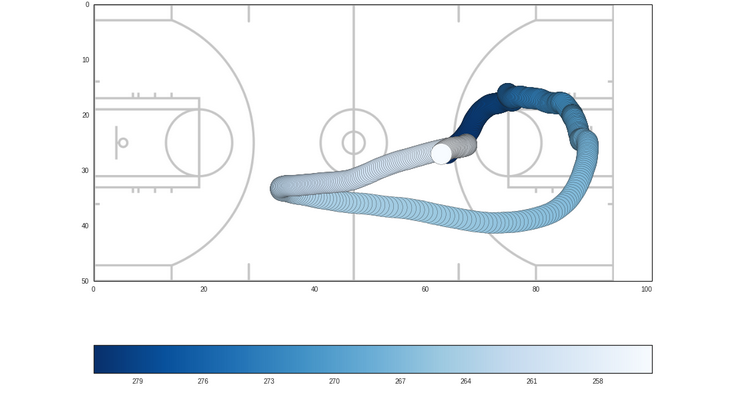
****

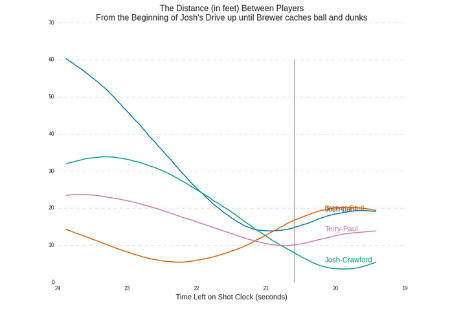
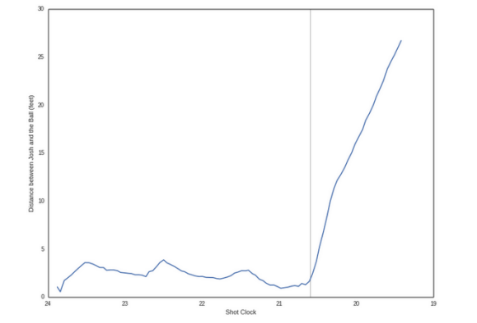
**Problem 2:**



**Problem 3:**



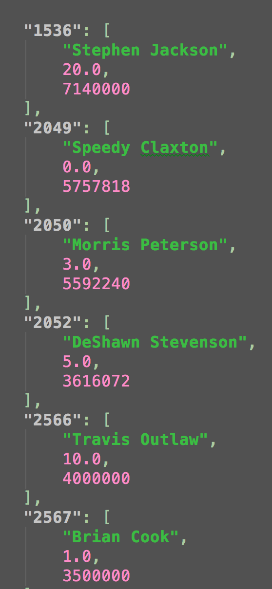




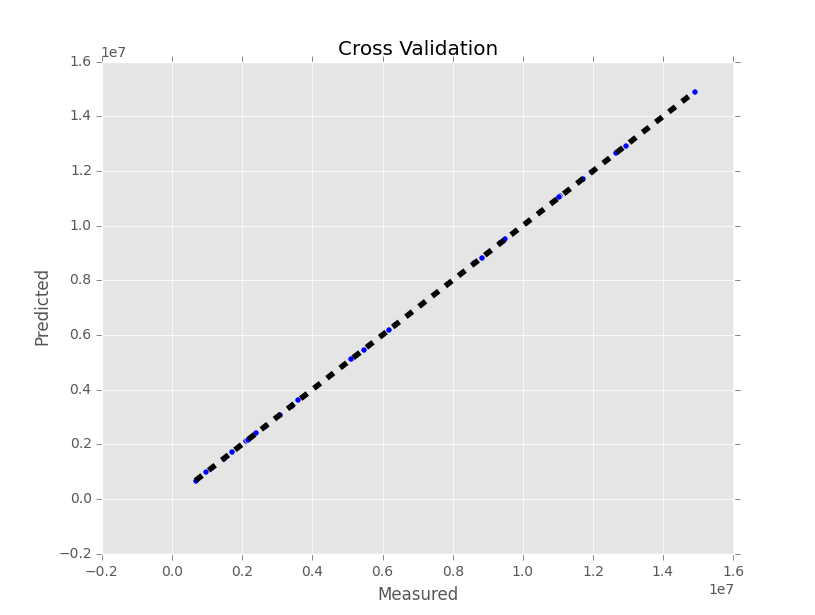
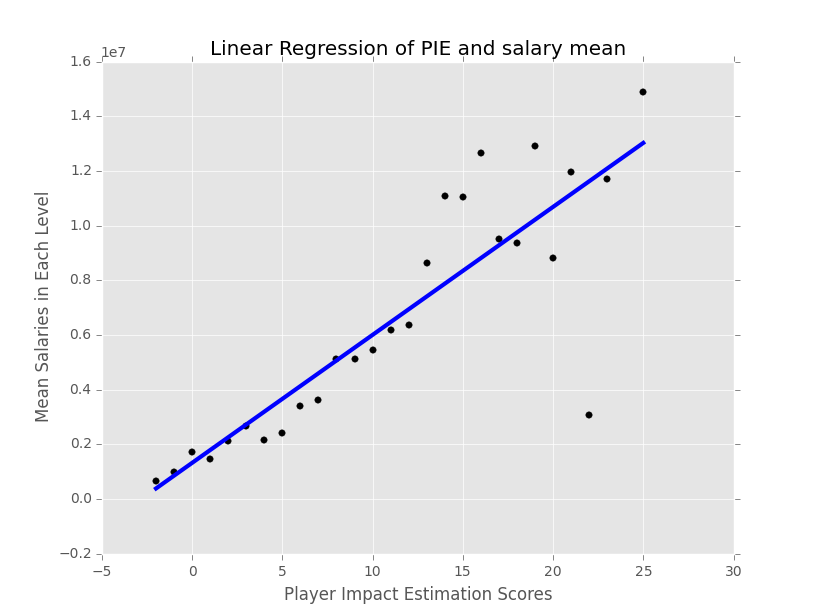
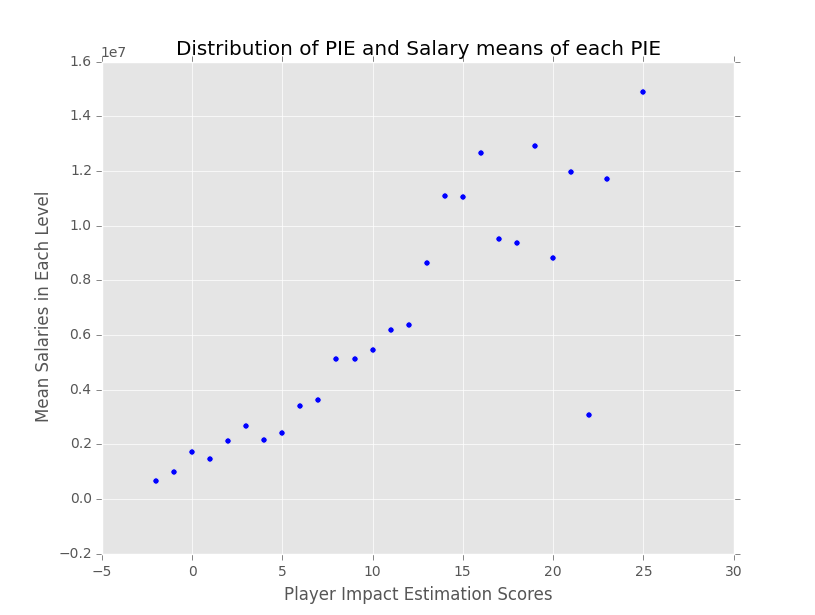
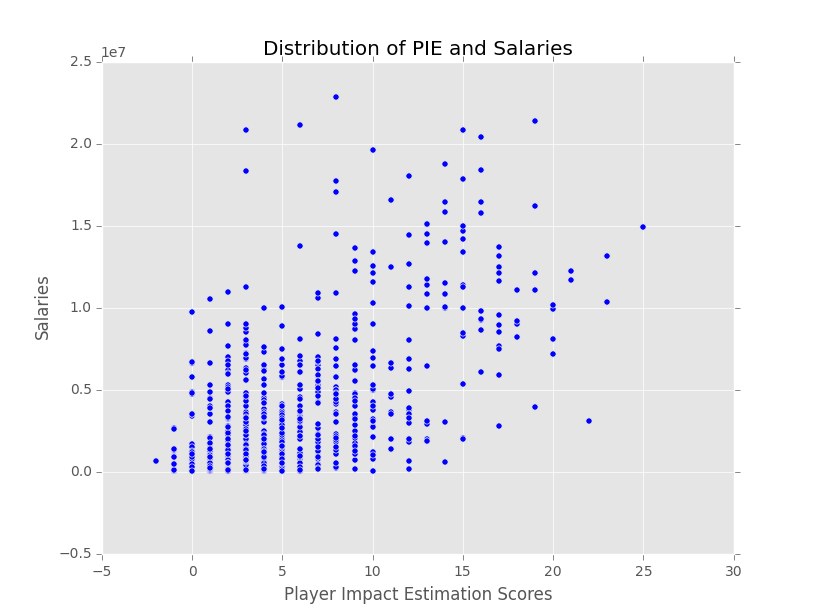
From this moment we can see an all-sided Josh Smith. He defended Griffin and took the rebound. He can dribble and have broad sight to make a great assist. He helped Houston a lot to win this important game and then won this series.

**Problem 4:**

1. NBA career stats data for each season



1. Data Distributions and Linear Regression



1. DEAs scores of each team

Los Angeles Clippers,1.418666667

Charlotte Hornets,1.086837253

New Orleans Pelicans,1.023390065

Golden State Warriors,1.015211346

Toronto Raptors,1.015009705

Portland Trail Blazers,1.011413521

Memphis Grizzlies,1.007647111

Atlanta Hawks,1.002109443

Dallas Mavericks,0.992895869

San Antonio Spurs,0.943466013

Boston Celtics,0.940914886

Chicago Bulls,0.931169202

Los Angeles Lakers,0.926424751

Utah Jazz,0.915267158

Brooklyn Nets,0.905647914

Detroit Pistons,0.902350726

Cleveland Cavaliers,0.899437081

Indiana Pacers,0.89798765

Houston Rockets,0.896929416

Milwaukee Bucks,0.896271972

Phoenix Suns,0.88451877

New York Knicks,0.883562756

Denver Nuggets,0.882659089

Oklahoma City Thunder,0.877791312

Sacramento Kings,0.877373649

Minnesota Timberwolves,0.877238508

Miami Heat,0.852223734

Philadelphia 76ers,0.847729932

Washington Wizards,0.840495103

Orlando Magic,0.839981093

**Conclusion**

1. Problem 1-3

In problem 1-3 we analyze player Josh Smith and the game 2015 playoffs HOU vs. LAC game 6. Josh is always a question player. How did he play in Houston? We can find some answers in that game. The series was 2-3 before that game, and in this game Clippers had a 19 points lead in 3rd quarter, which made everyone believe they will beat Houston and move on west final. However, Houston's benches took the control in the last quarter and completed an epic comeback. Josh Smith played a huge impact on this. In this moment, Houston succesfully defended Blake Griffen and Josh led a fast attack. So Houston tied the score and quickly surpassed later.

1. Problem 4

Experiment 1: Players’ Salary Based on Their Performance Impact Estimations

The first plot shows the general distribution of PIE and Salaries. These data points are highly biased so we couldn't conduct any regression to the original data points. However, we found that the salary mean of each PIE value lies on a line with PIE as shown on plot 2 and 3. The linear regression variance score is 0.74 which indicates it is a good fit. Plot 4 shows the cross validation result perfomed by sklearn library. We can draw the conclusion that PIE score has a linear relationship with salary mean. We can estimate a player's average salary that he will obtain by his PIE score.

Experiment 2: Predict NBA Final Champion in 2014-2015 season by DEA Model

Based on our DEA evaluation for team performance, the champion of 2014-2015 season should be Los Angeles Clippers. However, the real champion was Golden State Warriors. However, it has a high rank, 4th in our list. The opponent team of Golden State Warriors in NBA finals was Cleveland Cavaliers. It has 17th rank in our list, which implies Cleveland Cavaliers didn't perform really well in regular season competition. Predicting NBA finals is really interesting because it leaves a hard problem to tell which one is real the champion merely based on data. We think it is hard to predict it precisely because invisible human factors might have great impacts on the team performances.