NBA Players and Teams Analysis

Zibin Chen, Weifeng Lyu, Zeyu Ni, Jingya Qin, Jian Zhan, Yang Zhang
Georgia Institute of Technology

Introduction – Motivation

NBA inspires worldwide best basketball athletes to play for 30 teams in United States, attracts international audience in many countries to watch the games on site or via TV. Each NBA team plays 82 games, so there are 2460 games per season, which accumulates large volume of accurate dataset. Many data behind NBA has come to be scientific and valuable to analyze. Scouts, coaches and fans have their positive or negative attitudes to players and teams. The team hires data analysts to evaluate players and help them make trading decisions. Our project motivation is to deliver data presentation and analysis that convey meaningful implication and interesting analytic facts to the NBA fans through user interaction, maybe deliver profits by presenting valuable data visualization.

Problem definition

We decide to dive into performance of each player and each team, assign an overall score to each players, and each teams by means of Machine Learning analysis and data manipulation and evaluation to draw some interesting facts, which not only increase the readability and provide the overview of NBA data in user interface, but also provide support for champion prediction. Next, we use those scorings to predict the probability of each team to win the champion. The problem we define is to perform the analysis results of player level, team level and champion prediction level, with user interface such that the users can view their interested information.

Survey

Either evaluating a player or a team, or predicting NBA champion are not trivial, as many factors affect their performance. Scientific researches and predictions nowadays are well-conducted and can guide NBA team to make better decisions for trades, contracts, rookie selections [1] [2] [3]. However, most of the researches are conducted by the experts or scouts, which are private for themselves to generate profit for NBA teams. Those complex analysis and valuable researches are inaccessible and not understood from public. Our innovation focuses on integrating machine learning techniques with many existing complicated models to analyze statistics and implications in NBA research is rarely existed. Besides, NBA researches are usually conducted on the player level by scouts, while the quantitative evaluation of team is rare. Our targets are to establish the player rating system and complete the champion prediction system. The success is measured on how meaningful data is presented from the graph, and how effective our conclusion based on the back-test of our predictions with historical data. However, the algorithm is complicated, the project also aims to combine data analysis with data visualization. We develop an easy-understood way with a well-designed user interface demonstrate our analysis results with interactive user experience. The challenge of this project is significant amount of effort regarding to data preparation and result explanation that relies on NBA domain knowledge. NBA domain knowledge sharing, analytic skills and data analysis techniques along with this project experience are significant rewards for us as the payoff.

Proposed method

1. Intuition

The strength of our project delivery is demonstrated in usability and techniques. we propose the following intuitions:

(1). We use the most recent data. Many analysis and website user interface are using outdated data. The historical data we use include from 1999 to 2018 today.

(2). We follow official formulas and many

academic approaches as reference, such as PER, WS in Player Stats and Elo Score in team evaluation, which ensures our data authority.

(3). We integrate data from different sources, execute analysis towards those data and present solid conclusion. Not many projects can develop such board and in-depth analysis.

(4). We demonstrate good visual analytics. The board and deep analysis ought to be understood by people.

(5). We implement new scientific approaches such as SVM, KNN, Naive-Bayes, Decision Tree, Deep Neural Network, Time-series Analysis, Correlation Analysis, Cross Validation and Monte Carlo Simulation onto well-known dataset.

2. Description of your approaches

We aggregate the dataset from Kaggle to evaluate NBA players' statistics efficiency rating (PER) [4], which is the overall rating of a player's per-minute statistical production. Another classical measurement is Win-Share [5] [6] to determine how much a player contribute to the team's victory, comparing with the fact that PER is computed without considering game result. These two attributes are derived from many other NBA players' statistics such as scoring, minutes played and field goal percentages, to evaluate player's values in player evaluation module. As a team sport, the cooperation among players and the team chemistry are the important factor for winning the championship. We consider time-series influence for the total average scores for the different teams. Correlational analysis analyzes most significant factor for a team to win the game. After evaluating team, we are able to integrate some result to evaluate the Elo ratings, which is for calculating team's relative competitive levels in one on one situation and many researches are applied into NBA analysis scenario [7]. After the ratings of each team are determined, the probability of winning the final champion can be calculated by implementing machine learning techniques in decision tree, random forest with cross validation and Monte Carlo Simulation[8].

Beyond the evaluation steps, we also obtain the dataset from Kaggle, basketball-reference and photos from NBA Stats. These three websites store consolidated and updated data of NBA players and teams. We also utilize CanvasJS in JavaScript [9] and Python Flask Restful API along the project to present our analysis results to end-user. Users can then interact their input and check the ratings of their favorite players or teams. We also show the probability of each teams that could win the final champion.

a. Algorithms

(1). Player Evaluation

Some simple data merging, sorting and integration techniques run against the raw data and present visualization. SVM, KNN, Naive-Bayes, Decision Tree and Deep Neural Network with embedded features [10] are implemented for All-Star players prediction.

(2). Team Evaluation

Linear Regression model has been implemented to initiate a time-series analysis for all the teams' scores, which can remove the time trend bias for the following model. Decision Tree algorithm has been used to find the most important factors, which determines the outcomes of the games. Fed the algorithm with the data, Decision Tree model serves also as a predictive model to predict the winners of the champion and deliver the team evaluation outcomes. Team Elo rating is calculated based on past team game data following industry standard algorithm.

(3). Champion Evaluation

Autoencoder, Decision Tree, Random Forest, Monte Carlo simulation

b. User Interfaces

(1). Player Evaluation

The user interfaces are built with website HTML with Bootstrap templates, connected with Python Flask Restful-API. The end-user can browse player evaluation via website interface and interact with NBA player

seasons/playoffs data, most improved player and overall team Win-Share ratings, and run machine learning algorithm to predict All-star players by entering the input box and dropdown select

(2). Team Evaluation

The user interface is built with the website HTML and the data visualization model is based on the R Markdown and the ggplot2 packages.

(3). Champion Evaluation

The user interface is built with the website HTML and D3 visualization. A US map with each team and their winning probability is shown on the page.

Experiments/ Evaluation

1. Description of your testbed

a. Player Evaluation

a.1 How are the player stats changing over year?

a.2 Who will have the significant improvement next year and how this improvement contribution make sense?

a.3 How is team overall rating contributed from? How does it relate to different position players?

a.4 How are all-star selected and what are their characteristics?

b. Team Evaluation

b.1 How the scores develop with the time? This question will help to reduce the time-series bias when analyzing the teams

b.2 How the different stages of the seasons will influence the winner?

b.3 What is the correlation between the different important ratios, such as Elo-Score and the win rate?

b.4 What are the most important factors which will determine whether the team will win or lost a game?

c. Champion Evaluation

c.1 What is the winning probability a home team in a game?

c.2 What is the probability for teams to enter playoff?

c3. What is the probability for each team to win champion?

2. Details of the experiments

a. Player Evaluation

Player Evaluation included both visualization analysis and algorithmic computation. The final team score is based on the industry standard ELO rating mechanism and is used together with Player Stats and win share result in the final predictive model.

a1. Player Stats Module:

This module provides the end-user a glance of trend for players' total and per game data over their careers and their latest stat rankings among all other players by the choice of player name and game-series. This line chart implies some insight about the relation between player's career year and different statistics. Generally, the athlete reaches their peak in the middle of their career and strongly associated with their health. Win-Share weights both team wins and individual contributions, high WS indicates the team obtains high Elo-Rating probably.

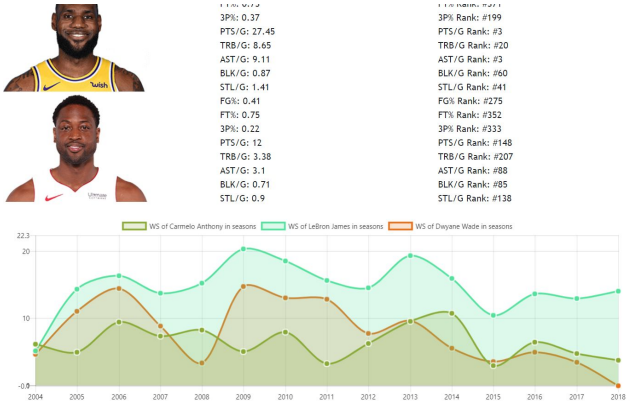

Figure 1. Player Win-Share Analysis

a2. Most Improved Player Module:

This module computes the values of score improvement derived by the formula, select the best 5 best scoring players and visualize their improvement. This bar chart captures the growth of the player to proven to be consistent with real NBA MIP selection. Normally, Those players are often in non-playoff team with low WS in previous year then lead the team to play well in the next season and make the playoff, the result from this module can pass additional factor for champion prediction to consider player improvement.

Figure 2. Most Improved Player Analysis

a3. Team Win-Share from individuals Module:

This module computes different positions of individual players' Win-Share and the sum of offensive Win Share and defensive Win Share for each team. We visualize the data in stacked bar graph intended to sum them up and evaluate the overall ratings of the team contributed by individual players. The height of the team bar reflects the winning probability of each team. Some teams with super-star player of one or more position, Win-Share rating is dominated by those positions. In these years, many teams with strong guards tend to achieve high overall Win-Share. The team Win-Share is considered as good attribute to further analyze in team evaluation and champion prediction.

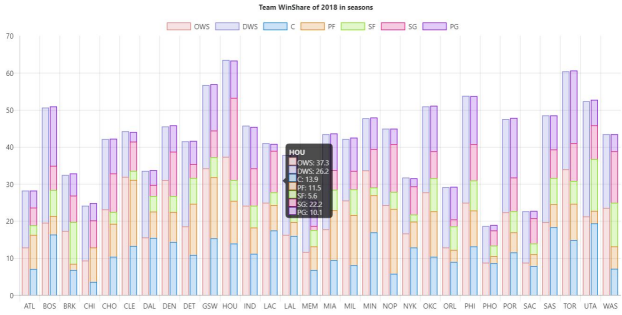

Figure 3. Team Win-Share Analysis

a4. All Star Prediction Module:

This module predicts the selections of all-star players based on many selected attributes in dataset, it combines supervised machine learning algorithms described above. The players predicted to be all-star will be shown in website interface and the end-user can visualize the player's percentile that exceeds how much percentage of players in the league. We select several most important attributes as inputs to run machine learning algorithms, the radar area of those selected players are large, meaning that all-star players are outstanding in those statistics. The result generated is convincing, those selected players often are voted to be all-stars in real-life. They have largest radar area, which validates the player with good game performance are expected to be voted as all-star.

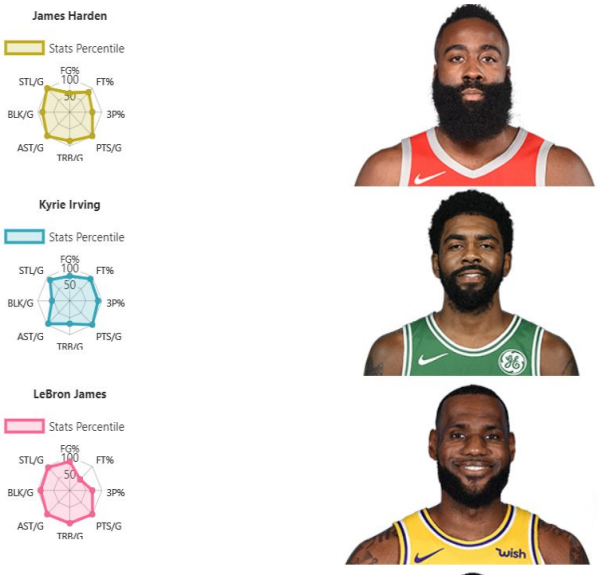

Figure 4. All Star Prediction Analysis

b. Team Evaluation

b.1,2 Time-series analysis:

The box plot has been used to analyze the time-series influence for the total average scores for the different teams, from the plot, it shows clearly that in the early years, the scores are relatively low compared to current years and the average scores keep fluctuating, but the trend is decreasing. This result is helpful for the further analysis about the score and the win rate.

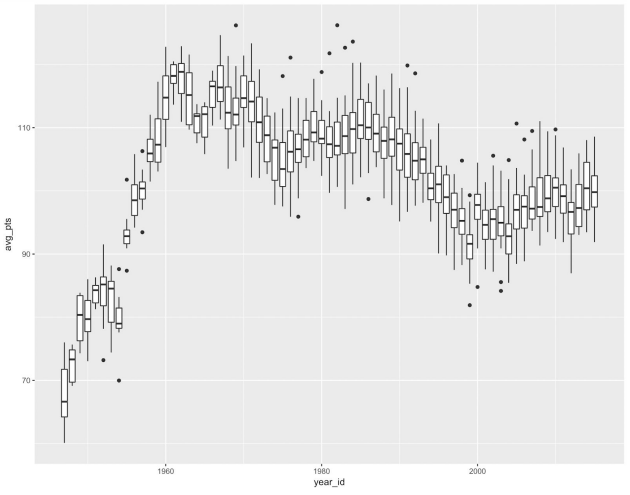

Figure 5. Box Plot for Total Average Scores

When the season can be categorized into different periods, it can also show some interesting patterns about the total average scores and the time. Usually the beginning of the season, teams tend to get a lower score, while in the middle of the season, the scores

increases, but decreases again in the end of the season. In some senses, it shows the team's strategy about winning the games, which is trying to win as many as possible games during the season and when approaching the play-off, they will save some players and prepare for the play-off, so the games will not be so aggressive.

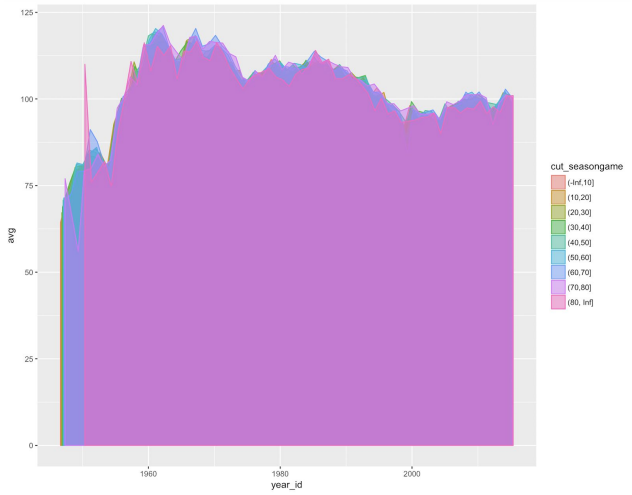

Figure 6. Scoring categorized in single season

b.3 Correlational analysis:

Correlational analysis is helpful for determining the relationship between different factors, in this part, firstly we standardize the data, based on the results from the previous analysis to remove the time bias and make all the data at the same scales. Then, the important factors have been implemented the correlational analysis and the figure is shown as the following.

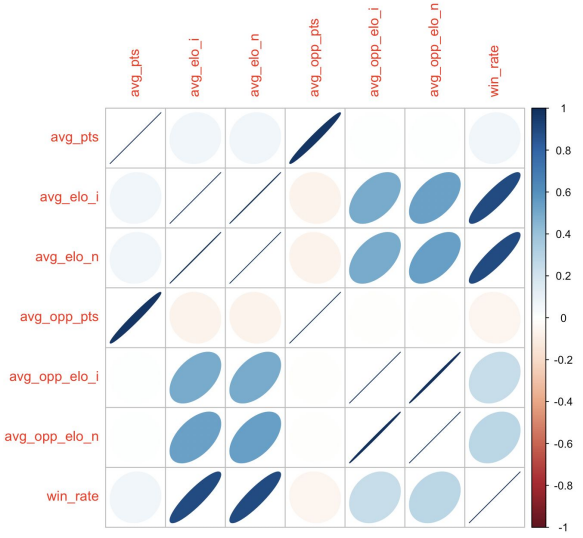

Figure 7. Correlational Analysis for Important Factors

There are some interesting observations:

(1). The win rate is more positively correlated with the self_elo, which means in some sense, whether the team can win a game is not decided by the opponent, but by themselves.

(2). The points is not very important, which means the high points cannot guarantee the team to win a game.

(3). The opponent's points are highly positively correlated with the team's points. It is interesting, because when combined with the previous point, it can conclude that the winning strategy is not to win as much as points, because in the same time, the opponent will also gain more points.

b.4 Important factors:

The correlational analysis is the qualitative understanding about the influential factors to the teams, when the decision tree algorithm has been used to create the predictive model, we can have a better understanding about how much and in what scale the factors will influence the teams and the winning games:

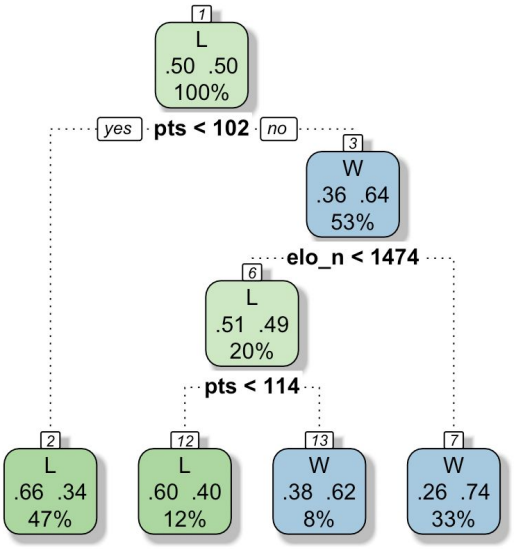

Figure 8. Predictive Model from Decision Tree algorithm

c. Champion Evaluation

c1: We trained the model in decision tree and random forest with cross validation to select the best model as well as the parameters. Then the model showed that the random forest with random forest will be the best model, which has a cross validation accuracy of 62.2%.

Below are the teams with the highest and lowest 5 winning probability. The result also makes

sense comparing with the NBA ranking now.

Home Team	Visitor Team	prob
Golden State Warriors	Phoenix Suns	0.807096
Golden State Warriors	Los Angeles Lakers	0.801811
Golden State Warriors	Brooklyn Nets	0.795354
Golden State Warriors	New York Knicks	0.788588
Golden State Warriors	Minnesota Timberwolves	0.782763

Home Team	Visitor Team	prob
Phoenix Suns	Golden State Warriors	0.333505
Brooklyn Nets	San Antonio Spurs	0.331828
New York Knicks	Golden State Warriors	0.327847
Orlando Magic	Golden State Warriors	0.306524
Brooklyn Nets	Golden State Warriors	0.305028

Figure 9. Highest and Lowest probability for winning the game

We also validated the model with ROC AUC score, precision, recall, and f1-score, the results are below. These results show the model works fine.

	precision	recall	f1-score	support
False	0.71	0.42	0.53	563
True	0.68	0.88	0.77	791
avg / total	0.69	0.69	0.67	1354

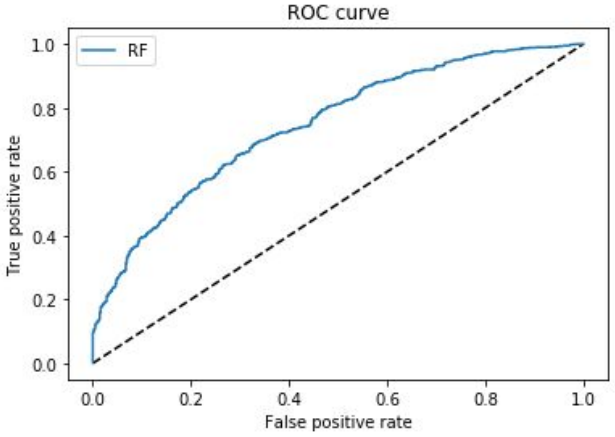

Figure 10. ROC Curve

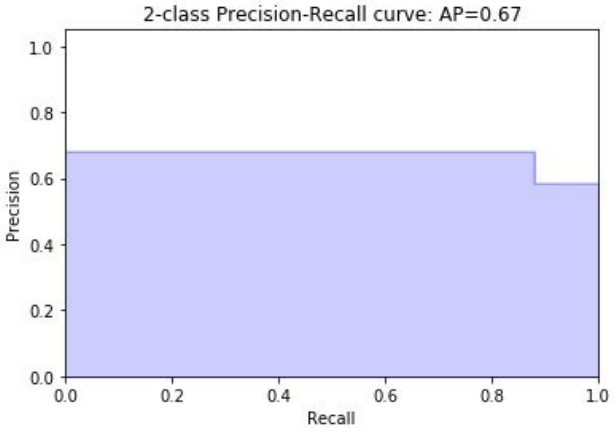

Figure 11. Prediction-Recall Curve

c2: After analyzing the historical data, we find the team with the highest ranking has the highest probability to enter the playoff round and they perform more like the geometric distribution as it decays very fast the ranking decrease. Therefore we determine to use geometric distribution to simulate it. The result for probability vs ranking is as below.

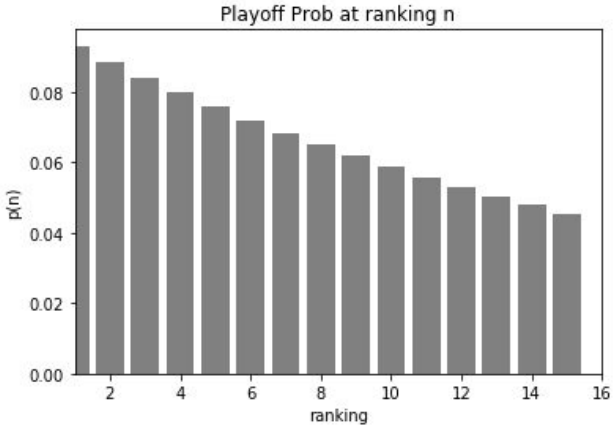

Figure 12. Playoff Probability at Ranking Analysis

c3: The probability of winning the champion depends on the probability to enter the playoff and probability of winning each round. We conducted the Monte Carlo simulation for 10 thousand times based on the winning probability(c1) and the entering probability(c2) to predict the final champion.

'Golden State Warriors': 2104,
'Houston Rockets': 1216,
'Toronto Raptors': 1152,
'Boston Celtics': 1079,
'San Antonio Spurs': 872,
'Cleveland Cavaliers': 582,

We also back-test it with the data last year to predict the champion last season.
'Golden State Warriors': 3508,
'San Antonio Spurs': 1742,
'Toronto Raptors': 896,
'Cleveland Cavaliers': 672,
'Oklahoma City Thunder': 527,
'Boston Celtics': 461,
The result for 2017-2018 season matched with the champion and the winner of eastern and western conference. Therefore, the validation shows that the model is well fitted.

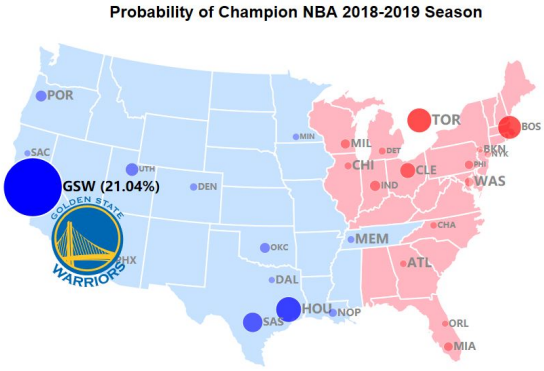

Figure 13. Final Champion Prediction Result

Conclusions and discussion

Final results make sense consistently because Win-Share analysis, all-star prediction, team analysis and champion analysis from different modules imply the champion for 2018-2019 is Golden State Warriors, which is the strongest team in the league now.

Our project is challenging and valuable though NBA data is highly available, it is one of the most popular sport topics in the world. However, we face a few difficulties while collecting those data that is ready for analysis.

(a). OpenRefine as the toolbox help us preprocess those null and convert some values into the types understood by the algorithm.

(b). We combine the raw data from multiple data sources by Python programming.

(c). We implement Python Scrape to pull those data.

(d). We study league rule change, player trade, team and player modification over years, the knowledge share helps to identify and resolve these issues.

We also encounter some learning curve to develop D3 and CanvasJS front-end, OpenRefine Data Cleaning, Python data integration and R, but all our team members contribute the strengths and learn from one another to consolidate the project perfectly.

Work Plan

Task	Team Member
Phase I: Preparation	
Task 1: Recruit Project Team Member	J.Zhan, Y.Zhang, Z.Chen
Task 2: Research topics and brainstorm methods	All members
Task 3: Manage and Organize Weekly Meeting	J.Zhan
Phase II: Proposal	
Task 1: Integrate thoughts into project planning	W.Lyu, J.Zhan
Task 2: Complete Proposal and Task Distribution	W.Lyu, J.Zhan
Task 3: Consolidate Video Recording and Presentation	J.Qin, Z.Chen
Phase III: Implementation	
Task 1: Data Collection and Integration (Python Scrape)	All members
Task 2: Players Evaluation (Python, sci-kit learn, Keras)	W.Lyu, Z.Chen,
Task 3: Teams Evaluation (Python, sci-kit learn, R)	J.Qin, Z.Ni
Task 4: Champion Prediction (Python, sci-kit learn, Simulation)	J.Zhan, Y.Zhang
Task 5: Result Visualization (JavaScript, Python, Jupyter Notebook)	All members
Phase IV: Summary	
Task 1: Progress Report	W.Lyu, J.Zhan
Task 2: Poster Design	J.Qin, Z.Chen

Task 3: Peer Grading	All members
Task 4: Final Report and Readme	All members

Distribution of team member effort

Zibin Chen	**Weifeng Lyu**	**Zeyu Ni**	**Jingya Qin**	**Jian Zhan**	**Yang Zhang**
16%	18%	16%	16%	18%	16%

REFERENCES:

[1]. S. Shea, Basketball Analytics: Objective and Efficient Strategies for Understanding How Teams Win, CreateSpace Independent Pub. Platform, 2013.
[2]. S. Shea, Basketball Analytics: Spatial Tracking, CreateSpace Independent Publishing Platform, 2014.
[3]. H. S. Bhat, L.-H. Huang and S. Rodriguez, "Learning Stochastic Models for Basketball Substitutions from Play-by-Play Data," University of California, Merced, 2015.
[4]. T. Zhang, J. Chen and X. Zhao, "Modeling and Analysis of Player Efficiency Rating for Different Positions: Case Study with NBA," in Advances in Computer Science and Education Applications, Beijing, Beijing University of Posts and Telecommunications, 2011, pp. 234-242.
[5]. Website, "NBA Win Shares," Basketball Reference, [Online]. Available: https://www.basketball-reference.com/about/ws.html. [Accessed 8 October 2018].
[6]. A. Olson-Swanson, "Predicting NBA Winning Percentage," 9 August 2018. [Online]. Available: https://towardsdatascience.com/predicting-nba-winning-percentage-in-upcoming-season-using-linear-regression-f8687d9c0418. [Accessed 10 October 2018].
[7]. V. Madhavan, "Predicting NBA Game Outcomes with Hidden Markov Models," University of California, Berkeley, 2015.
[8]. A. Groll, C. Ley and G. Schauberger, "Prediction of the FIFA World Cup 2018 – A random forest approach with an emphasis on estimated team ability parameters," Technische University Dortmund, Dortmund, 2018.
[9]. K. Dale, Data Visualization with Python and JavaScript: Scrape, Clean, Explore & Transform Your Data 1st Edition, Newton: O'Reilly Media, 2016.
[10]. J. Perricone and I. Shaw, "Predicting Results for Professional Basketball Using NBA API Data," Stanford University, 2015.