Unsupervised Learning Approach to Compete in Fantasy Leagues

Tanay Agarwal
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

Akshay Sethia
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

M. Indumathy
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India.

***Abstract*—In International sports and games, basketball is one of the most engaged sports across the world. Also, in recent years, the evolution of fantasy leagues have exercised the sporting fans of these games to show more enthusiasm and analyze the outcomes more intensely. In the world of basketball, the National Basketball Association (NBA) holds a special place in terms of fans, media, skills, and the value marked with each player. Throughout history, the volume of data and statistics that are collected each year for every player has increased drastically. With the availability of such enormous data, it is quite challenging to analyze and predict the outcomes of future games. The study adopts the unsupervised learning approach applying the K-Nearest Neighbors (KNN) algorithm to predict the statistics of the players in the forthcoming season. The projected statistics converted to fantasy points are used in the fantasy leagues. The comparison result confirms that the approach we have adopted display great ability in predicting the statistics regardless of the complexity in the data features.**

Keywords—Statistics, NBA, Fantasy Points, Similarity, Value-Based Drafting

I. Introduction

The National Basketball Association (NBA) has grown its attention worldwide and gathered millions of followers. Since the league is widespread, the coaches, fans, and specialists need to predict the outcomes of a match. Basketball players' statistics are described by various parameters such as points, assists, blocks, rebounds, turnovers.

Over the years, there has been an active growth in the popularity of fantasy leagues. It carries fans across the world to participate actively and create virtual teams comprising of real players. The players in fantasy leagues are defined with the help of fantasy points. Players with better statistics have higher fantasy points. These players then compete against other virtual teams created by other fans. In the end, the team with better cumulative fantasy points wins the league.

A large volume of data and statistics is collected in the NBA, which is increasing with every game played. Even with such rich data available, it is challenging to analyze and predict the outcome of a future match. Different machine learning methods have been implemented to deal with the complexity over the years on the data to simplify and gather significant knowledge.

The study adopts the use of an unsupervised learning approach applying the K-Nearest Neighbors (KNN) algorithm, where players are grouped based on their similar statistics. These grouped statistics used to project players' statistics for the forthcoming season. The fantasy points of the players are calculated based on their projected statistics. The model's efficiency is measured by comparing the projected data with real data. The results also show the model stands against other competing models

II. National Basketball Association

The NBA is a basketball league in which men basketeers take part and consists of 30 teams. Out of this, 29 teams are from the United States of America (USA), and the remaining franchise is vouched in by Canada. The NBA franchises partitioned into two conferences, which is the Eastern Conference and the Western Conference, and these contain three divisions, each of which is incorporated by five teams, which sum up the total to 30 franchises.

Each of the franchise needs to participate in 82 games in the regular season, partitioning into 41 matches at their arenas and 41 games at away arenas. Each franchise competes within their division, which counts to 16 games in 12 months. Each franchise has to play a total of 24 matches. In the end, all the franchises battle each other of the other conference two times a year, counting to 30 games in total.

TABLE I. DIVISIONS FROM NBA TEAMS

Western Conference			Eastern Conference		
Pacific	Northwest	Southwest	Central	Atlantic	Southeast
5 teams	5 teams	5 teams	5 teams	5 teams	5 teams

The NBA playoff season begins during the late April between the eight most excellent teams of both conferences. Those eight franchises are the victors from each division, a franchise that has the elite record of the conference after the division winners and the teams in the lower four segments. NBA championship playoffs are conducted in the tournament way. Each franchise plays in an organization of best of seven arrangement, and it implies that the franchise to get four wins can go ahead to the next round, on the other hand, the losing franchise will be ruled out of the competition. This process is rehashed in the accompanying rounds within the franchises in the same conference till only one franchise prevails. The final is held between the two franchises of each conference in the best of seven series. This series of events is known as the NBA finals, which is held customarily in June consistently.

III. Related Work

Numerous specialists have offered various methods in predicting the result of the games in the NBA, each attempting to present an accurate model to provide the best highlights of the game. The unpredictable nature made by the dream sports proffers a large functional play area for scholars to examine the issues of different methods like artificial intelligence, game theory, fundamental computer science and mathematics, game theory, and economics.

In contrast to the growing interest of fantasy leagues, studies concerning these games in scholarly writing are non-existent. Data and Information mining is employed to produce significant knowledge from data. Over here, classification performs an influential part. The task here is to analyze and group players with similar data across seasons to predict the stats of next season. In this manner, classification has a significant influence on the overall process. The efficiency of a model depends on how accurate its prediction matches with the real data.

In this situation, we are applying an unsupervised learning approach, which has no predefined titles for classification. Instead, the model distributes the data based on the attributes it selects during the primary stages. The paper "Evaluating

Machine Learning Varieties for NBA players' winning condition" gives information about the various methods that can be employed to calculate the performance of players and their impact on the winning percentage of their franchise. The paper presents techniques like Support Vector Machine (SVM), Random Forest, and Polynomial Regression to associate the players' performance. These methods provide consistent performance despite the complexity of data.

The main idea of the paper "Machine Learning Approaches to Competing in Fantasy Leagues for the NFL" by Jonathan R. Landers, Brian Duperrouzel, is on implementing a model that can help to select a team for the NFL (National Football League). Here the model included the use of Decision trees, Perceptron, Coalition Formation, and Multi-Agent Systems.

IV. Proposed Method

Statistics Projection of NBA players is a critical task when we have extensive data across all the seasons, and each player holding about 20 features detailing their games per season. It is challenging to select parameters that describe the potential of a player accurately. In the present work, the KNN algorithm is engaged to find players across ten seasons that have similar statistics and use that data to project the statistics of those players next season. This algorithm is quite popular among the researchers, and in many cases, this algorithm is quite successful in building a fitting model. The Root Mean Square Error method is employed To calculate the accuracy between our projected model and the real data

V. Experimental Methodology

The experiment carried out using Python and Jupyter notebook on the dataset retrieved from the official NBA statistic website using an API call. The following phases are implemented.

A. Collection of data:

An API call is made to the official NBA website to get the data. Another dataset taken from the official ESPN website is used for testing the efficiency of our model.

B. Cleaning and normalizing season data:

First, we import our data and drop the rows which have missing values. We also cut the player data who have played less than ten games to have an efficient model. For normalizing the data, every cell value is subtracted with the columns' minimum value. The value obtained is divided by the difference between the maximum and the minimum value of the column.

C. Creating a model:

Here, the idea is to compare and locate similar players across ten seasons. The distance formula is applied to calculate the similarity between the normalized statistics of two players. After the ten most similar players are found, a weight is assigned to these similar players by dividing there distance from 1. After attaching the weights to these similar players, these weights are multiplied to all the parameters of the data. Now to get the projected statistics, the values of each data column are added, and then the total value is divided by the total weight of the similar players.

D. Evaluating model:

To evaluate the model efficiency, we use the root mean square error. The error calculated between the projected data and the actual data helps in finding the confidence of the model. Also, we compare our model projections to the projections made by the official analytics team of ESPN.

VI. Algorithms

A. K-Nearest Neighbors (KNN):

K Nearest Neighbors (KNN) is one of the most simple, easy to apply algorithm which can work on both classification and regression problems. In this method, there is no training data available to the model. The model over here has to produce the appropriate output titles/classes for the data given to it. Here, we will be using classification past regression, as we will be either selecting or rejecting the player based on their statistics. KNN algorithm makes some preliminary assumptions to start with, which are similar points that lie within the same vicinity. Also, K over here defines the number of neighbors we will be choosing. KNN apprehends the idea of similarity, also known as distance, proximity, or closeness in technical terms. KNN calculates the distance between points and groups the points on that distance factor, taking into consideration the least distance between different locations to be as the center for that neighborhood. The speed of the algorithm is inversely proportional to the size of the data.

B. Root Mean Square Error (RMSE):

Root Mean Square Error (RMSE) or Root Mean Square Deviation (RMSD) is a method to estimate the deviation of the model from the desired output in foretelling data. It is represented by,

$$RMSE = \sqrt{\frac{\sum_1^n (x_e - x_o)^2}{n}}$$

Where, x_e *= expected value*

x_o *= obtained value*

n *= size of data*

Disregarding the division by n under the square root, the principal point is that this equation resembles the Euclidean distance equation between 2 vectors. So, RMSD can be considered as the separation between the vector of anticipated points to those of watched points. Over here, dividing by N assists in rescaling the Euclidean metric by a factor of √(1/n). If the value of RMSE is small, then the model is efficient in forecasting else vice versa.

C. Factor Adjusted Team Similarities (FATS):

Factor Adjusted Team Similarities (FATS) projection model stats that four elements of basketball - Shooting, Free Throws, Rebounding, and Turnovers provide a sketch of what affects the performance on the field.

Shooting is estimated by taking a look at the effective field-goal percentage (eFG%), which factors in additional points earned when making a hit from the bend/arc. Two groups with the same field-goal rate will contrast in eFG%, on the chance that one makes more three-pointer shots than the opposition.

Free Throws per field-goal Attempt (FT/FTA) displays how well a group shoots free throws, but not the attempts that it has taken. The best crew in this class is the one that gets lots of whistles, besides, convert their free throws to points.

Rebounding can be further classified as, offensive rebounding percentage (ORB%) and defensive rebounding percentage (DRB%), both of which show the level of potential sheets the crew being referred to gets on that individual side of the floor.

The last one is the Turnover Percentage (TOV%), which measures the number of turnovers a team records per 100 possessions.

Shooting, turnovers, rebounding, and free throws ought to be weighted 40 percent, 25 percent, 20 percent, and 15 percent, separately. Furthermore, that is for each side of the ball. There was no separation made between offensive rebound rate and defensive rebound rate. Hence, a fantasy point of a player is calculated by the formula, *Fantasy Points = Points + Field Goals Made + Free Throws Made – Field Goals Attempted – Free Throws Attempted + Offensive Rebound + Defensive Rebound – Turnovers + Assists + Steals + Blocks*

VII. Value-Based Drafting

The axiom of Value-Based Drafting is to discover players who are underestimated in rankings or average draft position, although their anticipated insights and yields ought to have them higher. So in simple words, Value-Based Drafting is a procedure that alters rakings of the player depending on his position. The thought behind Value-Based Drafting is to correctly rank players comparative with positions. Each position is just relative to a pattern. Some set the design to be substitution level or the last dynamic player drafted at each position. Others contrast every player with a mean or average at their particular positions. With the goal for Value-Based Drafting to be viable, everybody needs to rank players in precisely the same way. Not just that, they have to share the approach and ways of thinking. In a vacuum, they can anticipate a similar creation from a player. However, there is a whole other world to projection than that. There are unwavering quality, toughness, and consistency.

The player may share a bye-week or have an apparent troublesome matchup during their end of the season games. They may, as of now, have a player on a similar group on their list, and they want to enhance. The fact of the matter is, not exclusively will everybody have various desires for every player, except inborn potential to each dream crew vary. We might be verging on the instinctively self-evident, and that is ok. However, the suggestion with Value-Based Drafting is to continually take the top-positioned player on the board as they offer the most potential, at every determination, to the team. The objective is to put the same number of players that we rank most noteworthy on our list. Now and then, to achieve that, we have to go amiss from the request. Appropriate positioning is a science; powerful drafting is craftsmanship.

VIII. Results

The proposed method is applied against the data set, and the model is created. The sample output of the expected data and the actual data is shown in table II and table III, respectively. The error is calculated using the Root Mean Square Error between both the outputs and the error was found to be 1.97. Hence will conclude that the proposed method of using the unsupervised learning approach showed confidence in projecting the statistics by 98.03%.

TABLE II. SAMPLE EXPECTED DATA

player_id	pts	reb	ast	blk	stl	fg%	ft%	fg3m	min	tov
203932	16	7.4	3.7	0.7	0.7	0.45	0.75	1.6	33.8	2.1
201143	13.6	6.8	4.2	1.3	0.9	0.54	0.79	1.1	29	1.5
202329	9.4	7.5	1.3	0.4	0.8	0.44	0.9	1.2	28.3	0.9
202692	8.8	3.7	2	0.3	0.6	0.41	0.82	1	21.5	1
203518	5.3	1.6	0.6	0.2	0.5	0.35	1	1.3	19	0.5
203458	11.1	5.6	1.1	0.9	0.4	0.5	0.64	1	20.1	1.3
1627816	5.1	3.6	0.8	0.5	0.2	0.49	0.64	0.4	14.5	0.6

TABLE III. SAMPLE OBTAINED DATA

player_id	pts	reb	ast	blk	stl	fg%	ft%	fg3m	min	tov
203932	16.1	6.75	2	0.7	0.9	0.44	0.78	1.4	33.9	1.6
201143	12.3	5.8	3.2	1	0.7	0.47	0.76	1.1	30.2	1.6
202329	9.8	5.39	1.8	0.4	0.9	0.44	0.74	1.4	28	1.1
202692	8.3	2.42	1.5	0.1	0.5	0.41	0.77	0.7	20.8	1
203518	5.5	1.73	1	0.1	0.5	0.41	0.8	1.1	17	0.6
203458	8.3	6.87	0.7	1.3	0.5	0.51	0.65	0	23.3	1.2
1627816	2.2	0.93	0.5	0	0.1	0.45	1	0.3	9.1	0.3

We also compared our models' projected output to the official ESPN fantasy league projections. A sample output of the ESPN fantasy projection data and our models' projection data is shown in table IV and table V, respectively. The error obtained when we compare the models mentioned above is 2.11. Hence we can conclude that over models' projected data matched with the official ESPN fantasy league projected data with a confidence of 97.88%

TABLE IV. SAMPLE ESPN PROJECTION DATA

player_id	pts	reb	ast	blk	stl	fg%	ft%	fg3m	min	tov
203932	16	7.4	3.7	0.7	0.7	0.45	0.75	1.6	33.8	2.1
201143	13.6	6.8	4.2	1.3	0.9	0.54	0.79	1.1	29	1.5
202329	9.4	7.5	1.3	0.4	0.8	0.44	0.9	1.2	28.3	0.9
202692	8.8	3.7	2	0.3	0.6	0.41	0.82	1	21.5	1
203518	5.3	1.6	0.6	0.2	0.5	0.35	1	1.3	19	0.5
203458	11.1	5.6	1.1	0.9	0.4	0.5	0.64	1	20.1	1.3
1627816	5.1	3.6	0.8	0.5	0.2	0.49	0.64	0.4	14.5	0.6

TABLE V. SAMPLE PROPOSED SYSTEM PROJECTION DATA

player_id	pts	reb	ast	blk	stl	fg%	ft%	fg3m	min	tov
203932	16.06	6.75	1.99	0.71	0.91	0.44	0.78	1.4	33.94	1.58
201143	12.34	5.8	3.25	1.03	0.65	0.47	0.76	1.13	30.25	1.61
202329	9.78	5.39	1.75	0.45	0.95	0.44	0.74	1.41	27.98	1.14
202692	8.35	2.42	1.53	0.14	0.54	0.41	0.77	0.69	20.8	1.01
203518	5.5	1.73	0.99	0.11	0.48	0.41	0.8	1.07	17	0.57
203458	8.35	6.87	0.74	1.29	0.54	0.51	0.65	0	23.28	1.24
1627816	2.24	0.93	0.54	0.05	0.15	0.45	1	0.35	9.09	0.3

IX. Future Work

In summary, the unsupervised learning approach performed well in projecting are NBA players statistics. Our next step is to focus on rookie players as we do not have their current season data to predict anything. Additionally, future work will focus on getting results for players who have been traded which season and for players who are eligible for more than one position. We would also dive into other fantasy leagues for sports like football and cricket and try to build a model to win their respective fantasy leagues. Other than the significant steps, the future work also involves using different weights for the parameters involved in predicting the statistics.

References

[1] Evaluating Machine Learning Varieties for NBA Players′ Winning Contribution, Po-Han Hsu, SainzayaGalsanbadam, Jr-SyuYang and Chan-Yun Yang, Member, IEEE.

[2] Machine Learning Approaches to Competing in Fantasy Leagues for the NFL, Jonathan R. Landers, Brian Duperrouzel K. Elissa, "Title of paper if known," unpublished.

[3] A Machine Learning Based Approach Towards Building A Sustainability Model For NBA Players, SirshenduHore, Tanmay Bhattacharya.

[4] NFL Play Prediction, Brendan Teich, Roman Lutz, Valentin Kassarnig.

[5] Prediction of NBA games based on Machine Learning Methods, Renato AmorimTorres.

[6] NBA statistics website. https://stats.nba.com/

[7] ESPN NBA statistics website. https://www.espn.in/nba/stats

[8] Value-Based Drafting Fantasy Basketball website. https://socalledfantasyexperts.com/fantasy-basketball-value-based-drafting/

[9] ESPN fantasy basketball website. https://www.espn.in/fba/