NBA Statistics Data Analysis

NBA data analysis using data mining techniques

Nishant Batra
University of Alabama at Birmingham
Birmingham, USA
nishant@uab.edu

Ammar Khan
University of Alabama at Birmingham
Birmingham, USA
amk211@uab.edu

Rajeshvari Patra
University of Alabama at Birmingham
Birmingham, USA
parta@uab.edu

Janak Patel
University of Alabama at Birmingham
Birmingham, USA
janak@uab.edu

***Abstract*— We aim to detect outliers and outstanding basketball performers from the huge dataset. We use Machine Learning and Data Mining techniques to compare, analyses and visualize our datasets. Our ultimate goal is to draw conclusions on outstanding player groups, which is very useful for comparing and judging the performances of all players and figure out outstanding performers from the dataset.**

I. Introduction

In our work, we have worked with NBA 2004-05 dataset. We have implemented data mining techniques on the dataset for better understanding of the statistics. The three main actors in the dataset were Coaches, Players and Teams. All the analysis is based on either of them to find the top 20 of each type and then calculate their performance and visualize and cluster the data for better understanding and analysis. The outcome of this analysis can be very useful in deciding about the future players and coach to be chosen to build a strong team in the years draft.

II. Dataset Analysis

There were total eleven datasets from which two were for coaches, two were for teams and one was drafts and the remaining six were for players. The tree in Figure 1 can be an easy representation of all the data used for the analysis.

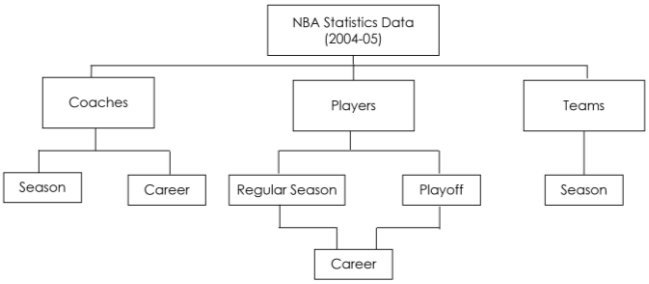

Fig. 1: Dataset decision tree

The datasets contained the records of season games, playoff games and even the entire career stats of the games. There were three datasets that contained all the basic information of teams, players and the drafts that took place. These were very useful for finding additional information about them.

A. Data Manipulation and analysis

Based on the statistics given, it was very difficult to analyze the data. So after researching more about the NBA games, some useful calculations were made. It was derived that the performance of a player can be calculated based on the positive and negative attributes of a player. So the formula used was:

[Points + Rebounds + Assists + Steal + Blocks + Field goals made] /

[Turnover + Failed Goals + Failed Attempts + Personal Fouls]

Where the first half contains the positive attributes and the second half contains the negative attributes of a player. The negative attributes of a player also needed some further calculations:

Failed goals = Field goals attempt – Field goals made

Failed Attempts = Free throws attempt – Free throws made

Performance was used as the key criteria for clustering and visualization of the datasets. The Coaches and the Teams dataset has win and loss records for season games and playoff games. So, average percentage wins depending on the win loss record could be calculated. All this calculation proved to be very useful for analysis of data.

B. Missing/Inconsistent data handling

There were some parts of the data that has a zero value, but they were still included in analysis. However, while calculating the performance, the main issue faced was that some entries had a divide by zero error. This was because either the players had inconsistent playing records, or the players had played very less number of games. So, these values were eliminated.

III. Prediction Using Decision Tree Algorithm

Using machine learning, prediction algorithm was implemented that predicts on the bases of a decision tree. According to our dataset, we have made predictions on the performance of top 10 players to get the average records. What basically was done was that this algorithm could tell us that whether a player's performance through the years is below average or above average. This can prove to be very helpful for future drafts and teams building strategy. Similarly, for coaches and teams, the main criteria were to find the top 10 coaches and teams that have above average performance in all the years they have played. The criteria that can be chosen for prediction can be

1. Entropy: The level of randomness
2. Gini: Impurity of the data.

For prediction algorithm, we calculated the man of performance/average wins so that our algorithm can

differentiate between underperforming and overperforming coaches.

Prediction was performed for the following datasets.

A. *Player Regular Season*

Figure 2 shows the top 10 players based on the performance. The first row shows the player id, second is the year of the stats, third and fourth are first and last names and fifth is the performance.

ilkid	year	firstname	lastname	performance
CHAMBWI01	1961	Wilt	Chamberlain	5657
CHAMBWI01	1962	Wilt	Chamberlain	5374
CHAMBWI01	1960	Wilt	Chamberlain	4722
CHAMBWI01	1963	Wilt	Chamberlain	4590
CHAMBWI01	1965	Wilt	Chamberlain	4518
ABDULKA01	1971	Kareem	Abdul-jabbar	4374
CHAMBWI01	1966	Wilt	Chamberlain	4331
CHAMBWI01	1967	Wilt	Chamberlain	4169
ABDULKA01	1975	Kareem	Abdul-jabbar	4147
CHAMBWI01	1959	Wilt	Chamberlain	4071

Figure 2: Top 20 players in season games.

We tried to predict the records for Wilt Chamberlain as it can be observed that he is an outstanding player because his name comes eight times in the top 10 list. The score of this dataset received was 0.80 from which it can be said that the dataset is complex, and the criteria used here was Gini. At the end, (1) was received as the output that means that the Will has performed above average in all the years he has played.

```
In [8]: model.fit(inputs_n, target)
Out[8]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                               max_features=None, max_leaf_nodes=None,
                               min_impurity_decrease=0.0, min_impurity_split=None,
                               min_samples_leaf=1, min_samples_split=2,
                               min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                               splitter='best')

In [9]: model.score(inputs_n, target)
Out[9]: 0.8053055671829217

In [10]: model.predict([[532,993,337]])
Out[10]: array([1], dtype=int64)
```

Figure 3: Output for prediction

B. *Player Allstar games.*

Figure 4 shows the top 10 players based on the performance.

ilkid	year	firstname	lastname	performance
ONEASH01	1999	Shaquille	O'neal	984
DUNCATI01	2002	Tim	Duncan	978
OLAJUHA01	1993	Hakeem	Olajuwon	970
OLAJUHA01	1994	Hakeem	Olajuwon	958
BIRDLA01	1983	Larry	Bird	941
BARKLCH01	1992	Charles	Barkley	935
JORDAMI01	1991	Michael	Jordan	912
BIRDLA01	1986	Larry	Bird	899
ABDULKA01	1973	Kareem	Abdul-jabbar	827
MCGINGE01	1974	George	Mcginnis	815

Figure 4: Top 10 Player in all-star games

The prediction performed for this table was for Shaquille O'neal. We got the output that this player has performed above average in all the years he has played.

C. *Coaches Season*

Here, we take the top 10 coaches having the best win records. The average win can be calculated as follows:

Average Win = ((Season Wins + Playoff Wins) / (Total Matches Played)) * 100

Where Total matches played = Season Wins + Season Loss + Playoff Wins + playoff Loss.

coachid	year	firstname	lastname	AVG_Wins
CUNNIBI01	1982	Billy	Cunningham	85.7879925
JACKSPH01	1995	Phil	Jackson	85.5691057
COSTELA01	1970	Larry	Costello	83.1010453
DALYCH01	1988	Chuck	Daly	82.5322812
JONESKC01	1985	KC	Jones	82.5203252
SHARMBI01	1971	Bill	Sharman	82.0731707
JACKSPH01	1996	Phil	Jackson	81.5468549
JACKSPH01	1990	Phil	Jackson	81.312769
RILEYPA01	1986	Pat	Riley	81.300813
POPOVGR01	1998	Gregg	Popovich	81.1176471

Figure 5: Top 10 coaches in season games

It can be observed that Billy Cunningham comes on the first place with average win percentage of 85.78. When the algorithm was carried out for this dataset on Billy, it was shown that he has above average wins.

Just to trick our algorithm, we tried with coach whose average wins comes on the 510 rank.

510	SLOANJE01	2001	Jerry	Sloan	39.3292683
511	HARRIDE01	1979	Del	Harris	39.2857143
512	FITZSCO01	1978	Cotton	Fitzsimmons	39.2682927
513	MCMAHJA01	1968	Jack	McMahon	39.2276423
514	DUNLEMI01	1991	Mike	Dunleavy	38.7195122
515	FRATEMI01	1994	Mike	Fratello	38.7195122
516	RIVERDO01	2000	Doc	Rivers	38.7195122

Figure 6: 510-516 ranking of coaches

After performing this prediction, the output received was (0), that means that the coach has all the average win records below average.

D. *Teams season*

The team season dataset was straight forward except for one thing. The Team names were given in shortcut form so for locating information of any team, we needed to refer the team's dataset.

team	year	Average Wins
CHI	1995	87.80487805
LAL	1971	84.14634146
CHI	1996	84.14634146
PHI	1966	83.95061728
BOS	1972	82.92682927
BOS	1985	81.70731707
CHI	1991	81.70731707
LAL	1999	81.70731707
WSC	1946	81.66666667
KEN	1971	80.95238095

Figure 7: Top 20 teams in season games

As observed in figure 7, the team that comes on the number one rank is CHI. Upon checking the team's dataset, the information of this team can be obtained.

CHI	Chicago	Bulls

Figure 8: Team information form team dataset.

As seen in the figure above, the team's name is Chicago bulls.

IV. K-Means Algorithm

K-Means clustering is one of the simplest and most popular unsupervised machine learning algorithms, used now a days for cluster analysis and Data Mining processes. We used this algorithm to generate clusters for our data set as well. In our scenario, python **jupyter** IDE was used as a working directory for the K-Means cluster generation and analysis. Here we used all those NBA data set which are directly related to players. Hence, we also used the same data set to solve similar classification problems as well. The problem domain considered here was total games played and performance. Our core objective was to find the number of group of players and check how homogenous these player groups are. To generate K-means cluster, the attributes which were used are total games played and player performance. But before generating K-means, we had to find the optimal numbers of clusters to be generated. Hence, elbow method was used here to find the optimal numbers of clusters that should be generated.

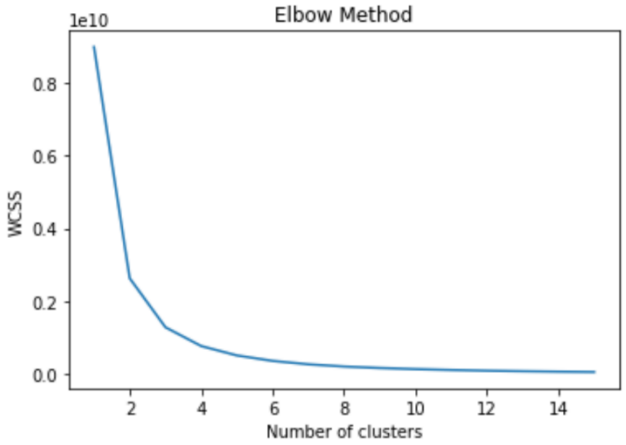

Figure 9: Elbow method

Above figure shows the implementation of elbow method in our scenario, here a graph is generated with a curve or a certain angle on a point. The angle lying on a certain point shows the number of optimal clusters to be generated. In our scenario, the optimal number of clusters are four. As a result, K-means was generated with four different colors clusters showing grouping of all those players with homogeneity. K-means also generated yellow points in every cluster. These points are called centroids of each cluster. The centroids are used to make the cluster visualization more understandable. The K means cluster generated shows four different clusters.

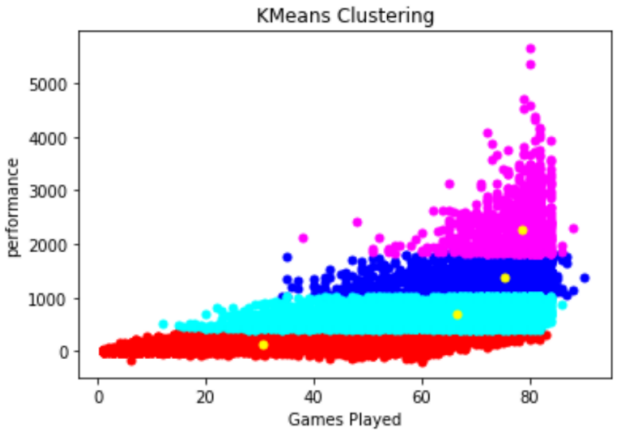

Figure 10: K-means on Player season games

Our first cluster for player regular season as shown above shows four clusters of different player groups. The first red cluster at the bottom shows some players ranging from not playing a single game to playing up to 80 games max. But their performance was below average, considering their performance as very poor despite of playing too many games as compared to other players. In the second blue cluster, there are players who played from 15 to max 80 games but their performance was somehow better than the players grouped in cluster number 1. In the third dark blue cluster, players had an average performance with games played from 40 to 80. And in the final pink cluster, players with the best performance were grouped. In this cluster there are some players who played up to 40 games but have an average performance as compared to players grouped in cluster 1 who played 80 games with poor performance. In the final cluster, there are players who played the most games and have the best performance out of all the players in the data set.

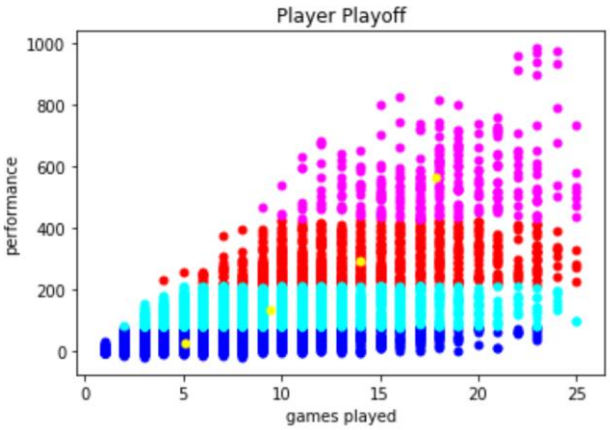

Figure 11: K-Means on player playoff games

Our Second Cluster for player plays off as shown above explains four clusters of different player groups. The first blue cluster at the bottom shows some player ranging from not playing a single game to playing up to 23 games max. But their performance was below average, considering their performance as very poor despite of playing too many games as compared to other players. In the second blue cluster, there are players who played from 2 to max 25 games but their performance was somehow better than the players grouped in cluster number 1. In the third red cluster,

players had an average performance with games played from 4 to 25. And their average performance was between 200 to 400. And in the final pink cluster, players with the best performance were grouped. In this cluster there are some players who played up to 25 games but have an average performance as compared to players grouped in cluster 1 and cluster 2 who played up-to 25 games with poor performance. In the final cluster, there are players who played the most games and have the best performance out of all the players in the data set.

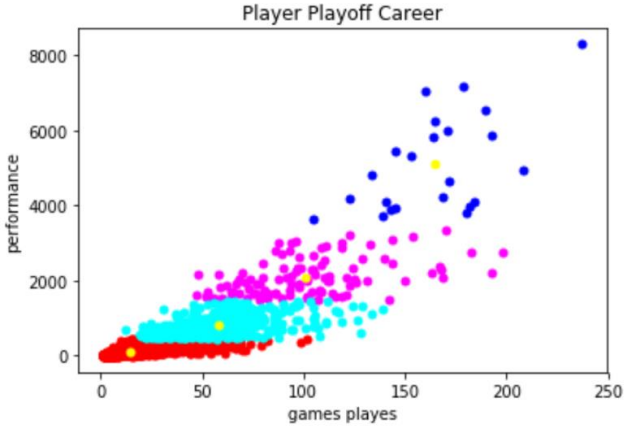

Figure 12: K-Means on Player playoff career games

Our third cluster for player playoff career as shown above shows four clusters of different player groups. The first red cluster at the bottom shows some players ranging from not playing a single game to playing up to 100 games max. But their performance was below average, considering their performance as very poor despite of playing too many games as compared to other players. In the second blue cluster, there are players who played from 20 to up-to 150 games but their performance was somehow better than the players grouped in cluster number 1. their average performance was between 500 to 1200. In the third pink cluster, players had an average performance with games played from 50 to 210. And their performance was between 1200 to 3500. And in the blue cluster, players with the best performance were grouped. In this cluster there are some players who played up to 240 games but have an average performance as compared to players grouped in cluster 1 and cluster 2 who played 140 games with poor performance. In the final cluster, there are players who played the most games and have the best performance out of all the players in the data set. And their performance was between 5000 to 8200 max.

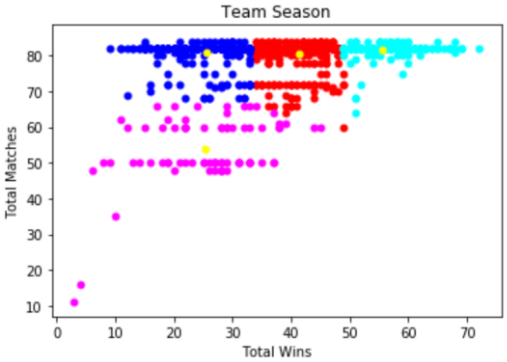

Figure 13: K-Means on Team Season games

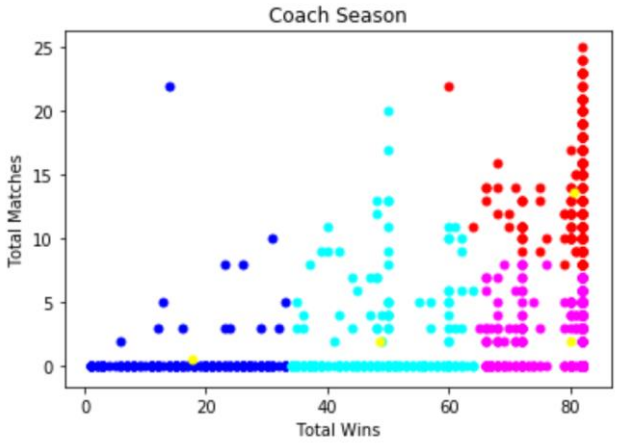

Figure 14: K-Means Clustering on Coach season games

V. DBSCAN ALGORITHM

Density-based spatial clustering of application with noise (DBScan) is a well-known data clustering algorithm. Which is used widely in data mining and machine learning. Based on sets of points DBScan groups together points that are close to each other based on a distance measurement including Euclidean distance and a minimum number of points. It also marks as outliers the points that are in low-density regions.

For, measuring DB-Scan algorithm we basically require 2 parameters. First, eps: The minimum distance between two points. It means that if the distance between two points is lower or equal to this value (eps), these points are considered neighbors. Second, Min-points: The minimum number of points to form a dense region.

Algorithm 1 The pseudo code of the proposed technique DMDBSCAN to find suitable Epsi for each level of density in data set	
Purpose	*To find suitable values of Eps*
Input	*Data set of size n*
Output	*Eps for each varied density*
Procedure	*1 for i* *2 for j = 1 to n* *3 d(i,j) ← find distance (x_i, x_j)* *4 find minimum values of distances to nearest 3* *5 end for* *6 end for* *7 sort distances ascending and plot to find each value* *8 Eps corresponds to critical change in curves*

Figure 15: Pseudo code for eps value

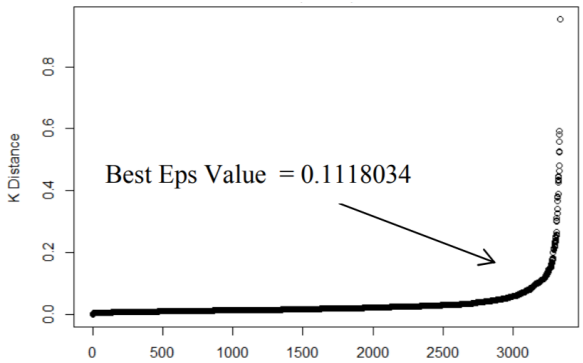

Figure 16: Best Eps value

In our scenario we used DBScan to find outliers for our datasets. For, DBScan first we have to find best eps(E) value. We cannot put random eps(E) value because it entirely depends on properties of datasets to datasets. For DBScan there is procedure to follow for getting best eps

value by taking one single point on our datasets and from that point we measure distance from all the other points. Then we move to second point and we do same for 2nd, 3rd, 4th data and we take all the distances. Eps value to be chosen is small, which means large part of dataset will not be clustered. If we choose high value, clusters will merge, and majority of objects will be in same datasets. For our particular dataset we choose one datapoint and then we have list of distances for all data points. Here, we are suggesting 3 nearest neighbors of that particular data point. We do it for each of the datapoints, we keep constructing the k distance with index of object. Whenever we find curve in critical change that will be best eps value. For that datasets. So, we implemented this concept in our scenario. In our case the best eps value is eps=0.5.

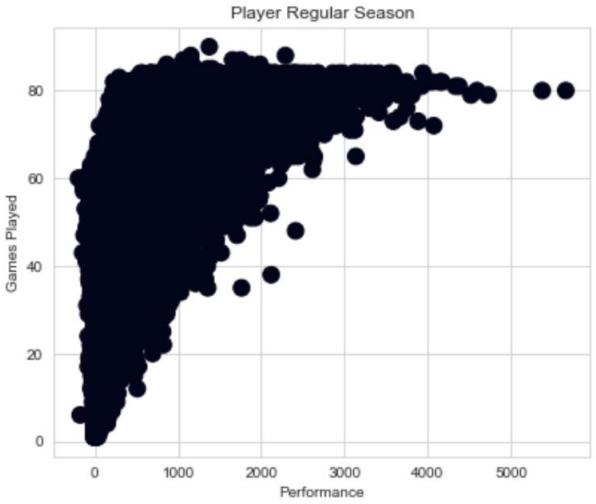

Figure 17: DBScan on Player season games

Above figure player regular season shows players as games played 0 to above 80 games and respect to performance is above 5000. But, some players did not play more games, but their performance is very good and above the average.

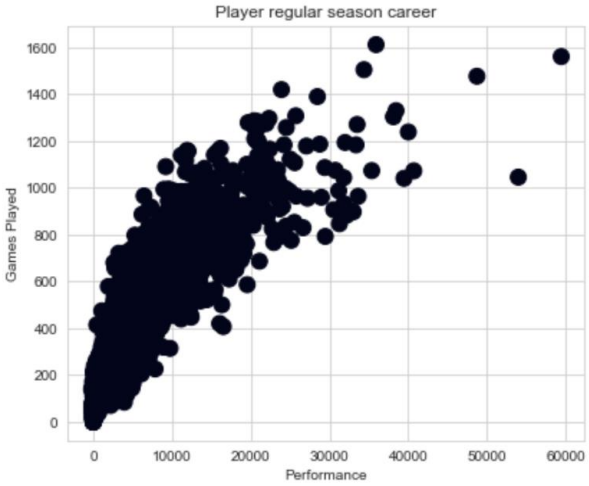

Figure 18: DBScan on Player season career games

VI. VISUALIZATION

In this Project, we did calculate players best performances and did figure out the outstanding players based on the performances. After figuring the top performances from huge data set, we further continued data analysis on attributes, because these are keys for the calculation of performance calculation. So, finally we did further analysis on, which attributes contribution played key role on performance in this NBA data set. The trend is useful for players to increase their performance of players, in further games.

To complete the above task, we have taken top performance list of four cases, namely player regular season, player regular season career, player playoff ad player playoff career for analysis. After analyzing the given data attributes for basketball players, we made a conclusion that the dataset we analyzed having six positive affecting attributes and four negative affecting attributes. So, the positive affecting attributes are points, rebounds, assists, steals, blocks, field goals. And negatively affecting attributes are missed field goals, missed free throws, turnover and personal fouls.

The first part of the attributes effects analysis is about on positive ones. For better visualization we opted line graphs. When we plotted the line graphs of top performances on y-axis and six positive attributes on x-axis, we can easily understand which attribute effects more on performances. The following are the pictures of positive affecting attributes.

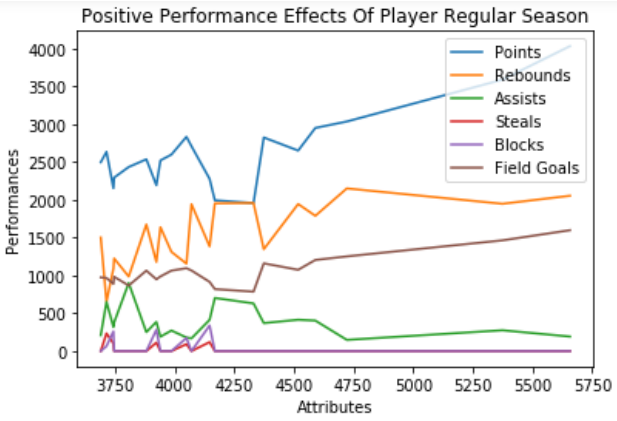

Figure 19: Line graph showing positive performance effect on player regular season games

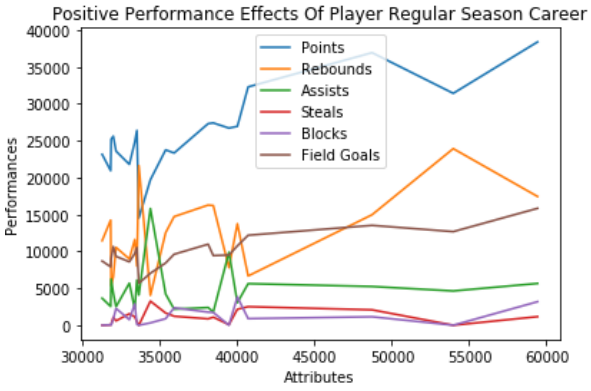

Figure 20: Line graph showing positive performance effect on player regular season career games

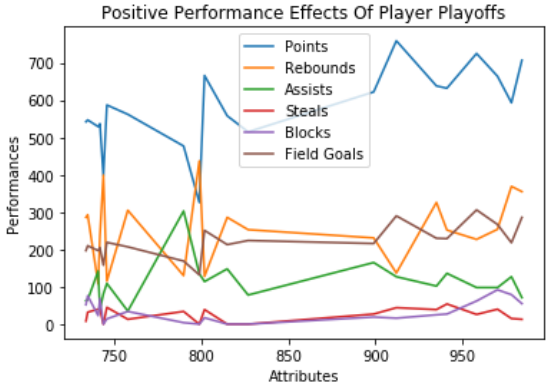

Figure 21: Line graph showing positive performance effect on player playoff games

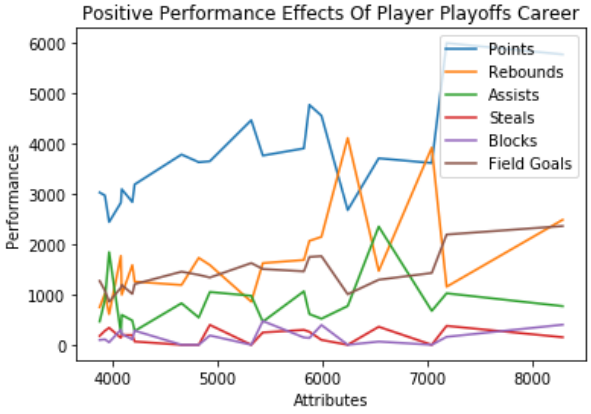

Figure 22: Line graph showing positive performance effect on player playoff career games

After plotting the line graphs of six attributes in four cases, we can see top lines effecting proportion on performance is more and bottom one is least effecting attributes. And we can visualize that the most effecting attribute is points and least effecting attribute is blocks in all four cases, even though some fluctuations are there. So, for the best players the most portion of their performances is their points, and least portion of their performance is their blocks. Finally, the positive attributes effects positive on performances, the more the number the better the performance of player.

The second part of the attributes effects analysis is about on negative ones. For better visualization we opted line graphs. When we plotted the line graphs of top performances on y-axis and four negative attributes on y-axis, we can easily understand which attribute effects more on performances. The following are the pictures of negative effecting attributes.

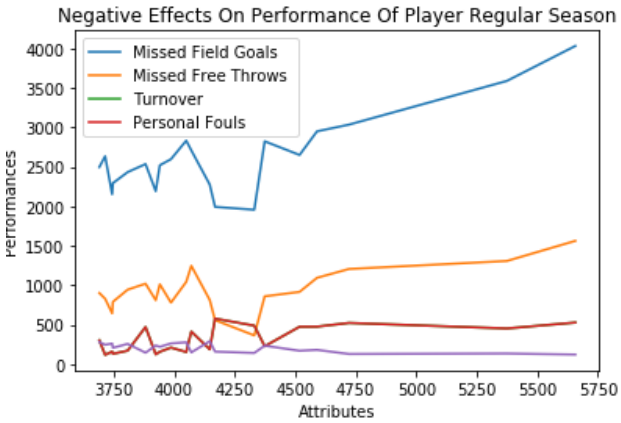

Figure 23: Line graph showing negative performance effect on player regular season games

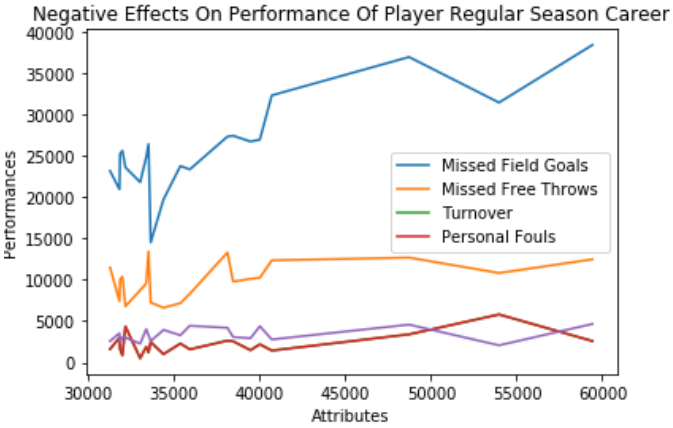

Figure 24: Line graph showing negative performance effect on player regular season career games

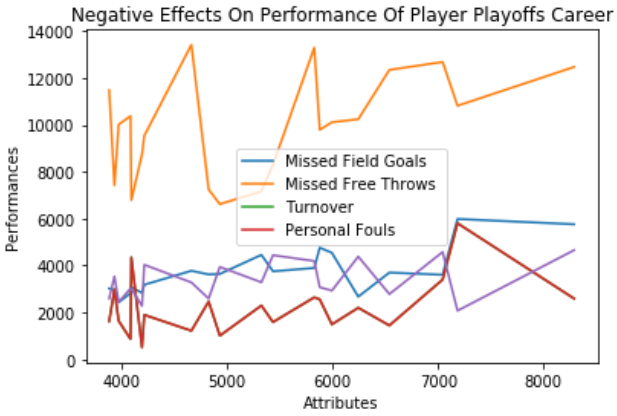

Figure 25: Line graph showing negative performance effect on player playoffs career games

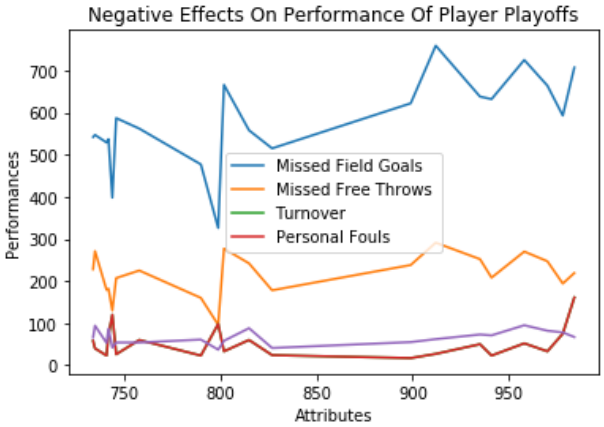

Figure 26: Line graph showing negative performance effect on player playoff games

After plotting the line graphs of four attributes in four cases, we can see top lines affecting proportion on performance is more and bottom one is least affecting attributes. And we can visualize that the most effecting attribute is missed field goals in first three cases and last one is missed free throws and least affecting attribute is turnover in all four cases, even though some fluctuations are there. So, for the best players the most negative affecting attribute of their performances is their missed field goals and missed free throws, and least negative affecting attribute of their performance is their turnover. Finally, the negative attributes effects negative on performances, the more the number the reduction on performance of player is also more. Finally, these analyses are very useful to understand the player performance effects and their proportion for period. Which is automatically helpful to the player for the future plays.

VII. Conclusion

We have done the analysis on NBA statistic data. We did draw conclusions on outstanding player groups, which is very useful for selecting players for future games. If the team is good obviously the success rate will increase. We were implemented of data mining techniques in our project like K-means and DB-scan. These algorithms are very useful for large data set implementations. Using K-means algorithms we did visualize the cluster as four in our data set. These four clusters could be best, good, average and poor performers. There also few people who played less games, but their performance is more. We can encourage them to play more games if there is medium competition in play. If there is highly competitive game, it's better to suggest players who played more games as well as more performances. Like this we can draw conclusions based on the K-means algorithm clusters in perspective of basketball game. And one another important advantage of clustering is very helpful for determining their salaries. If the player has more experience and best performance, we can say he is eligible for good pay. As the performance and experience increases their chances of increasing their earning also more. Using DB-scan algorithm we can see the outlier performers in the play. Outliers include both outstanding performer and very poor performers. And using line graphs we can further understand more on attributes of basketball players and attributes effects on their performances. These analysis and conclusions are very much useful for understanding how the basketball game wins and listing top performers which is very informative, and results should be easily understand by common man.

References

[1] YouTube, 16-Nov-2018. [Online]. Available: https://www.youtube.com/watch?v=PHxYNGo8NcI&t=770s.

[2] Codebasics, "codebasics/py," GitHub. [Online]. Available: https://github.com/codebasics/py/blob/master/ML/9_decision_tree/9_decision_tree.ipynb.

[3] "National Basketball Association," Wikipedia, 16-Apr-2019. [Online]. Available: https://en.wikipedia.org/wiki/National_Basketball_Association.

[4] "Basketball statistics," *Wikipedia*, 10-Nov-2018. [Online]. Available: https://en.wikipedia.org/wiki/Basketball_statistics.

[5] "Matplotlib Line chart," *python tutorials*. [Online]. Available: https://pythonspot.com/matplotlib-line-chart/.

[6] "sklearn.cluster.KMeans¶," *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[7] "A demo of K-Means clustering on the handwritten digits data¶," *scikit*. [Online]. Available: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html.

[8] M. J. Garbade and M. J. Garbade, "Understanding K-means Clustering in Machine Learning," *Towards Data Science*, 12-Sep-2018. [Online]. Available: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

[9] "How to normalize data to 0-1 range?," *Cross Validated*. [Online]. Available: https://stats.stackexchange.com/questions/70801/how-to-normalize-data-to-0-1-range.

[10] Nadia Rahmah and Imas Sukaesih Sitanggang 2016 IOP Conf. Ser.: Earth Environ. Sci. 31 01 2012. Available: https://iopscience.iop.org/article/10.1088/1755-1315/31/1/012012/pdf