Casestudy 4 Report

Team member: Shaowei Gong, Houjue Wang, Wenjing Li, William Grudzinski

1. Business proposition

As the commercialization and the expansion of NBA. The NBA is become more and more popular and gain bigger business market .The revenue involved in Professional Sports is increasing rapidly and the best way to see that is through the contracts of the athletes. In 2012, David Stern said that the NBA predicts the revenue of the league to be $5 billion in the 2011-2012 season, which would be a 20 percent increase from the last season. Now the league is worth $37.5 billion dollars and is growing rapidly.

It is obvious that the performance of the NBA team is vital to thrive it in the business market and improve it’s value along with the profit. So the performance is directly affect the revenue of the NBA team along with the salary structure of the whole team. Below chart is the team value of golden states warrior in the past five years:

Here is the performance of warrior in regular season in the past five years:

As we can see in the two charts, performance directly determine the value of the team.

2. Motivations

The team management want improve the team value which ensure a bright future of the team really need to get better performance in the NBA battles against other team competitor. As data scientists, we created a algorithm to analyze the every player’s scoring habit and the cluster of the all the player in the NBA league based on their performance statistic like scoring habit, assist ,rebound ,dense, speed and steal ect, so that we can persuade the team management to buy our idea and implementation to improve the team performances.

For example, by know the players scoring habit of the competitor, the coach of the team can make effective defense tactic to restrict the key player performance of competitor; also by knowing the scoring habit of players in their team, the training coach giving suggestion to get much more effective ways to score and make the right offensive tactic based the scoring habit of the players in their team and may change some player shooting habit to gain a better balance in the whole team.

Also, by using algorithm , the management of the team also can know about the player of the teams fall into the which cluster like dynamic SG/PG/SF/PF/C and slow SG/PG/SF/PF/C. The coach of the team can make the best or most appropriate tactics , offensive and defensive style based on the player cluster. Also, on the other hand ,management can choosing the players based on the clustering map to building a team according to the coach’s style of running the team; the salary structure of the team, because there lot’s of players fall in the same cluster, the management can buy the player in a specific cluster who has relatively low salary if the team has moderate financial power; the trade tactic, sometimes there too much some kinds of players fall in the same cluster that the team doesn’t need that much kind of players, they can trade some of the player they don’t really need to get the specific kind of player in the trade marketing to utilize the resources most. In conclusion, there lots of ways to mining the players’ performance data, deep understanding the how the data effect the players and even the team. Thus the management can improve the team performance by utilize our algorithm.

3. Methodology for analyzing the data?

In this prototype of our product, we provide 2 functionality that help NBA team analyze data. First functionality focus on analyzing shooting choice of players. Second functionality focus on cluster players.

To implement the first functionality, we divide the half basketball court into 30\*30 small section evenly. And we count the number of shootings within each small sections and calculate the shooting average field-goal percentage. The core of our analysis highly depends on data visualization. We use size of a circle to represent the number of shooting in each section and use the tone of color to represent the shooting average field-goal percentage.

To implement the second functionality, first, we clean and reorganize our data retrieved from API. We set the PlayerID as the index in Pandas dataframe, and then to make sure our analysis is valid we filter out players which do not play more than 41 games or the average playing time is lower than 15 minutes. And then merge various data retrieved from API according to PlayerID. So that we get a large set of feature that measure the performance and the gaming style of players.

In the next step, our goal is to cluster our players and display them in the way that can be easily understood by our client. To visualize our result of clustering, we need to reduce the dimension of our dataset to 2 dimensions, so that we visualize our data in common way. Thus, we calculate the correlation matrix of our dataset, which is core of almost every dimension reduce method. Then we use truncated SVD to process our correlation matrix and reduce the dimension to 10 dimensions. Note that singular value decomposition breaks any matrix X down so that , where , ,  and  are the singular values of the matrix X with rank r. Truncated SVD performs linear dimensionality reduction by setting all but the first k largest singular values equal to zero and using only the first k columns of U and V. It is very similar to PCA, but more flexible since it operates on any matrix directly, instead of merely on a covariance matrix. In this case study, we chose k to be 10 since it gave us a relatively close approximation to X. We do this, because it will give us more potential for our further development of our product, for example truncated SVD can be applied directly to dataset, which also can effectively reduce the dimensions of our data.

Then, we used t-SNE (t-Distributed Stochastic Neighbor Embedding) to visualize our data in a 2-dimensional space. The Python package is sklearn.manifold.TSNE.

t-SNE is a tool to visualize high-dimensional data. It converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. The advantage of t-SNE is that it is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales. It also produces significantly better visualizations than other existing techniques by reducing the tendency to crowd points together in the center of the map.

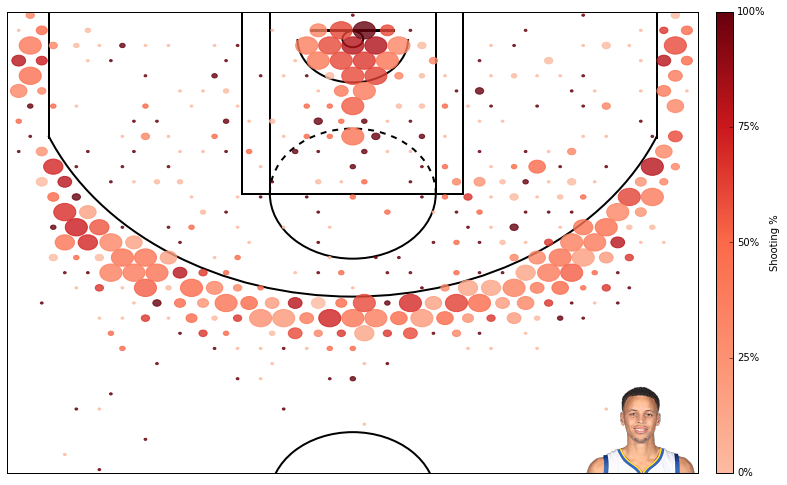
Finally, we apply K-means clustering to our dimension-reduced data. We iterate k from 1 to 15, and use wss as our standard to chose the best K. The last problem for us is how to display our clustering result to our client. We use scatterD3 library in R to visualize our data, which and create an interactive scatter plot.

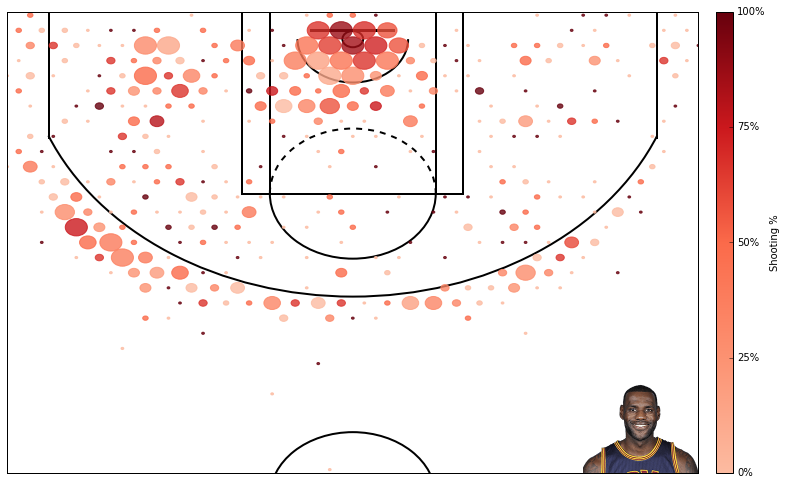
4.Result display

Please see the pictures below,

In this example, we show 2 shooting choices chart based on data of NBA 2015-16 regular season stats. The first one shows the shooting habit of Steven Curry, the second one shows the shooting habit of Lebron James.

These two chart provide the coach a lot of information. For example, I am the coach of Boston Celtics, and we will play against Cleveland Chevalier. Our biggest problem is how to defense Lebron James. Look at shooting chart, we can intuitively notice that Lebron’s shootings concentrate on hoop area. He does not shoot a lot in outside and the average field-goal percentage is relatively low in 3 point line and middle distance. Thus, we should invest our defense attention against Lebron in hoop area and left rib area in 3 line.





Originally, the chart below is an interactive graph in HTML format. But the word cannot embed HTML file into it, so blow is a screen shot. We have put the example HTML file into folder.

This diagram also is also very helpful for coach. For example, we need a strategy that can effectively limit Steven Curry, but there is no successful experience before. We find that Darrian Lillard is very similar to Steven Curry, we happen to have tactics that successfully limit Darrian Lillard before, then we can refer these experience.

