CISC 5950 Bigdata Programming Project 1

Suwen Gu, Qike Yan, Jiayuan Tong

Question 2: NBA Shot Logs Data Analysis Report

2.1 Introduction:

Background:

In the NBA, a top player makes around a thousand shots during the entire regular season. So as you can image, this data summarizes every shot made by each player during the games in the 14/15 regular season along with a variety of features, could be the real big data. We'll dive into the shot log, and focused on analyzing who is the most unwanted defender for each pair of players(A,B) known as the highest fear sore and the comfortable zone of some popular players' shooting. In terms of big data, we'll set up data on google cloud based on HDFS and develop a MapReduce- based algorithm to compute and classify our targets.

Data:

This is the Data on shots taken during the 2014-2015 season. Each row represents a shot and the columns are the details of the shot. There are 21 columns included who took the shot, where on the floor was the shot taken from, who was the nearest defender, how far away was the nearest defender, time on the shot clock, and much more. Here, the first target is to find the highest fear sore for each player. So, we target at miss and hit and player name and defender name, those four columns. For the second target about the comfortable zone of shooting is a matrix of these three columns: SHOT_DIST, CLOSE_DEF_DIST and SHOT_CLOCK. As the target analyzing players, we choose James Harden, Chris Paul, Stephen Curry and Lebron James.

2.2 Design:

Sub Question 1:

The target of this question is for find the most unwanted defender of each player based on the highest fear sore we calculated. The fear score defined as the hit rate of player A when facing other players. The mathematical formal is missed shot divided by total shots taken. So the core MapReduce design of this part is: mapper is responsible for mapping all needed data for each player and reducer is for computing the fear sore based on each specific defender. The final result will be the top one from sorted list which is the fear sore based on each possible pair of (player one, player two)

Sub Question 2:

The purpose of this question if for find the most comfortable zone for specific four players. What we need is using K-means as the classifier to find 4 zone (cluster) for each player and then we are using the similar method state above as the question one to find which zone could contribute the highest hit rate. Since it will involve multiple MapReduce for different computing based on the principle of K-means. The overall framework could be briefly described as following:

K-means:

As the question defined, we need to divide data (a sub dataset of original dataset based on the specific player from the four players list) into four zones. So, the first step is randomly

picking four data points as the initial centroids. Then for each data point in that sub dataset, we assign it to one centroids based on the closet distance between four centroids. After this, we will update each centroids based on new cluster data obtained above. What we'll update is assign the overall average to it. So far, we have already finished one circle of this classifying process. What we need to continue is repeat the steps as one loop until the difference between each old centroids and updated centroids is very small, which is the terminal condition.

MapReduce:

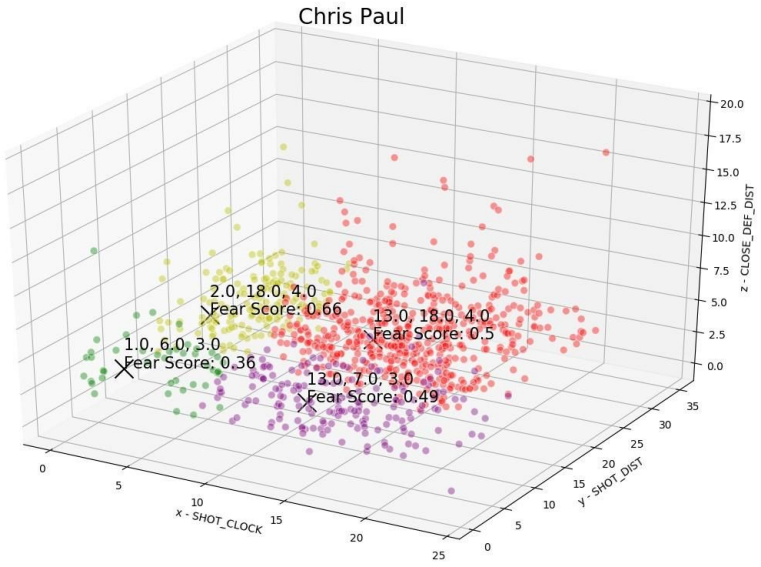
This question will use multiple MapReduce processes since the whole process could be divided into four individual tasks. 1. For filtering original dataset as four sub datasets based on each specific player name. 2. For randomly picking four data points as the original centroids for each sub dataset. 3. For running the K-means algorithm in the iterative way with loop unit meet the terminal condition. 4. For writing the final centroids for each data point as another column feature. As the final result is to find the most comfortable zone base on four clusters, we still need one more MapReduce to do find the highest hit rate among these four clusters for each player.

2.3 Results

Sub Question 1:

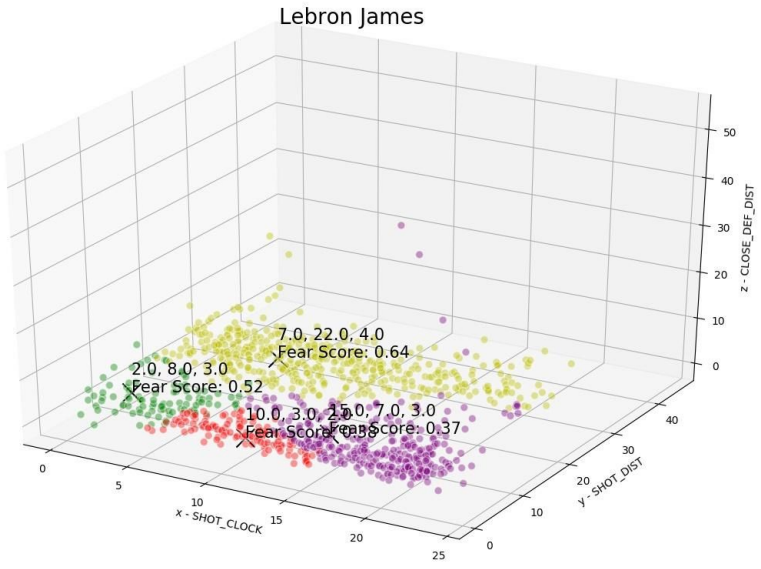
Please see appendix 2.

Sub Question 2:

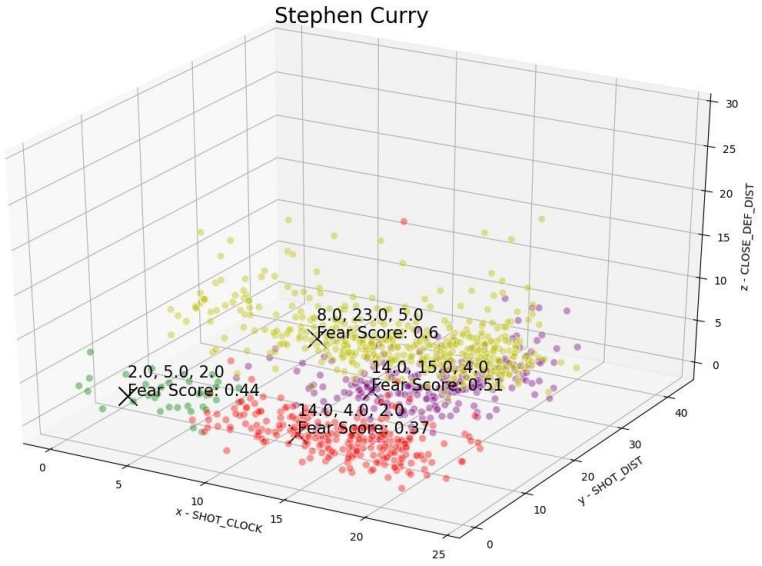


For Chris Paul, his most effective zone is when shot clock is 1s, shot distance is 6 feet away from the hoop and the close defender distance is 3 feet away.

For Lebron James, his most comfortable zone is when shot clock is 15s, shot distance is 7 feet away from the hoop and the close defender distance is 3 feet away.



For Stephen Curry, his most effective zone is when shot clock is 14s, shot distance is 4 feet away from the hoop and the close defender distance is 2 feet away.



From the graph below, we can see that James Harden takes a lot of shots. When the shot clock is 14s, shot distance is 5 feet away from the hoop, and he is encountering close defense, he hs the highest performance.