DataScience Final Report – Another Iverson

A071547 鄭澄遠

A061502 陳文揚

A061618 黃有德

A. Motivation

以往過去在五打五的全場籃球比賽中，場上各位球員的進攻以及防守位置，我們通常是依照各位球員的身材來進行區別，然而這樣的分類方式，或許也同時限制了各個球員的發展，因為還要考慮到每位球員打球的習慣，還有本身的強項在哪個領域，或許某位身材高大的中鋒，他的傳球視野並不輸給現役的任何一位控球後衛，也許讓他來擔任進攻傳導的角色，能使他在場上的效益更大。

然而在場上的比賽過程中，我們通常注意的數據，不外乎是：投籃命中率、抄截、籃板...等等，然而關於更細部的資訊，例如：持球時間、運球數、最近防守人距離，甚至是防守者是誰，或許更左右著這個球員在場上的數據，若是我們可以把這些資訊考慮到，也許對於這些球員有更客觀的強弱評斷。

這次的 Final project，我們嘗試將 NBA 球員的一些細部的資訊，納入分類的參考依據，並且做了 weighting 的計算，來重新將現役的 NBA 的進攻\防守能力進行排序，也許可以發現某些並非明星球員，但是也有相當亮眼的成績，並且再將這些數據進行分群，希望可以透過球員平常打球的習慣，可以重新歸納他們在場上打進攻/防守位置，也許也能夠透過這樣的數據，找出更多不同的搭配以及戰術的規劃。

B. Problem Formulation

我們的 data 總共有 21 一個 feature，每一次進攻對應到一筆 feature：

a. GAME_ID : 每個比賽對決的 index
b. MATCHUP : 紀錄比賽的時間及哪兩隊對抗
c. LOCATION : Home (H) or Away (A)
d. W : W 代表勝利，L 代表失敗
e. FINAL_MARGIN : 兩隊比分差距
f. SHOT_NUMBER : 第幾次投籃
g. PERIOD : 第幾節
h. GAME_CLOCK : 這個 play 的時間標記，用 mm:ss 表示(00:00~12:00)
i. SHOT_CLOCK : 投籃前，籃板顯示剩餘的進攻時數
j. DRIBBLES : 投籃前運球數
k. TOUCH_TIME : 持球時間
l. SHOT_DIST : 投籃距離
m. PTS_TYPE : 得分為兩分球或三分球
n. SHOT_RESULT : made 代表投進，missed 代表沒投進
o. CLOSEST_DEFENDER : 最近防守者名子
p. CLOSEST_DEFENDER_PLAYER_ID : 最近防守人的 index
q. CLOSE_DEF_DIST : 最近防守人距離
r. FGM : 有投進為 1，沒投進為 0
s. PTS : 得分
t. player_name : 紀錄當下 play 的 player 名子
u. player_id : 代表 player index

C. Proposed Model

我們基於上個部分所介紹的二十一個參數為基礎，且分成兩個面向：進攻以及防守，來對每個球員的能力做分析。我們分別定義了數個參數，再用這幾個參數當作 feature 當作 input 來分析各個球員，以下分成進攻、防守以及分類三個部分作介紹：

a. 進攻：

進攻的部分我以上面其中幾個參數作為基礎定義了幾個參數：

I. 命中率(Field goal percentage, FG%)：這項是比較傳統的數據，計算方式為投球的次數分之投球成功的次數，此項數據越高代表球員嘗試進攻的成功率越高，是分析球員進攻能力最直觀的方式。

II. 得分(Points, PTS)：這項也是傳統的數據，計算方式為該球員在此球季總得分為多少（不計入罰球），得分越高，贏得比賽的機會就越大。

III. 進攻難度(Difficulty of Offense)：這項是定義的新參數，計算方式為在進攻成功的前提下 FGM/CLOSEST_DEF_DIST，是投籃成功與否乘上投籃距離除以最近防守者的距離，針對每次成功的進攻計算一次之後再取平均。此數據的意義為投籃成功的前提下，完成此次投籃的難度為何，在這邊我們假設投籃的難度與投籃的距離成正比，且與防守者的距離成反比。

IV. 進攻效率(Efficiency of Offense)：這項也是我基於 dataset 的數據定義的新參數，計算方式為在進攻成功的前提下 PTS/TOUCH_TIME，是用來判斷進攻效率的參數。進攻效率越高代表在越少的時間內可以得到分數，對球隊的獲勝有很大的幫助。

b. 防守：

防守的部分定義了幾項參數當成判斷的依據：

I. 防守率：將所有防守過球員投進的次數除以總防守的次數當成防守率，這也是最直觀的方式看出一個防守球員有多少能力可以阻擋進攻方得分

II. 防守距離：資料為每一個 play 防守者距離進攻者投出球時的距離，把每一個 play 防守距離加起來除以總防守次數當成防守球員平均防守距離，這邊的直覺是，防的距離如果夠貼近越能給進攻者壓力，代表這位球員的防守能力也越強。

III. 持球時間：所有被防守的進攻者持球時間相加除以總防守次數，取這項 feature 原因是我們認為如果防守強度越強，那麼進攻者越不易甩開防守者製造空檔投藍，所以持球時間也會變久。

IV. 運球次數：所有被防守的進攻者運球次數相加除以總防守次數，這項 feature 與前面持球時間有相同意思，如果防守者越強的話，進攻者相對運球次數也會較多。

c. 分類：

分類的部分我們定義了幾個綜合的參數當成判斷的依據：

I. 平均投籃次數：總投球次數/總出賽場次

II. 平均投籃距離：總投籃距離/總進攻次數

III. 命中率：進球次數/總投球次數

IV. 持球時間：總持球時間/總進攻次數

V. 2 分/3 分球比：(2/3 分球次數)/總投球次數

VI. 運球次數：總運球次數/總進攻次數

D. Experimental Results

a. 排序進攻與防守者的強弱：

想法：我們想利用前面所算出的進攻/防守者的 feature，分析出每個球員進攻及防守強弱的資訊，方式則是利用 Borda count 的方式進行球員之間的大小排序，根據每個 feature，我們可以定義出對應的 ranking，例如以命中率為例，player A 為 50%，player B 為 70%，player C 為 40%，那我們可以求出在命中率這個 feature 下，三位球員的排名為 B>A>C，以此類推，有多少個 feature 就有多少個排序的 list，而我們的目標就是把這些 ranking list 疊加起來，產生出一組 consensus ranking list。

b. 做法：假設有 N 個 feature set，M 為球員，因此我們可以產生出 N 個 ranking list，每個 ranking list 中有 M 個元素(M 為球員的排序)，如果球員的排序在第一順位，就　　給予分數 M，第二位則為 M-1，以此類推，最後一名則為 1 分，因此每位球員會得到 N 個分數(根據在 N 個 ranking list 中的排名)，把這 N 個得到的分數相加當成是球員的最後總得分，在依據總得分排名可以得到 final ranking，我們便可以得知每位球員各自在進攻及防守上的表現。

I. Offensive ranking result

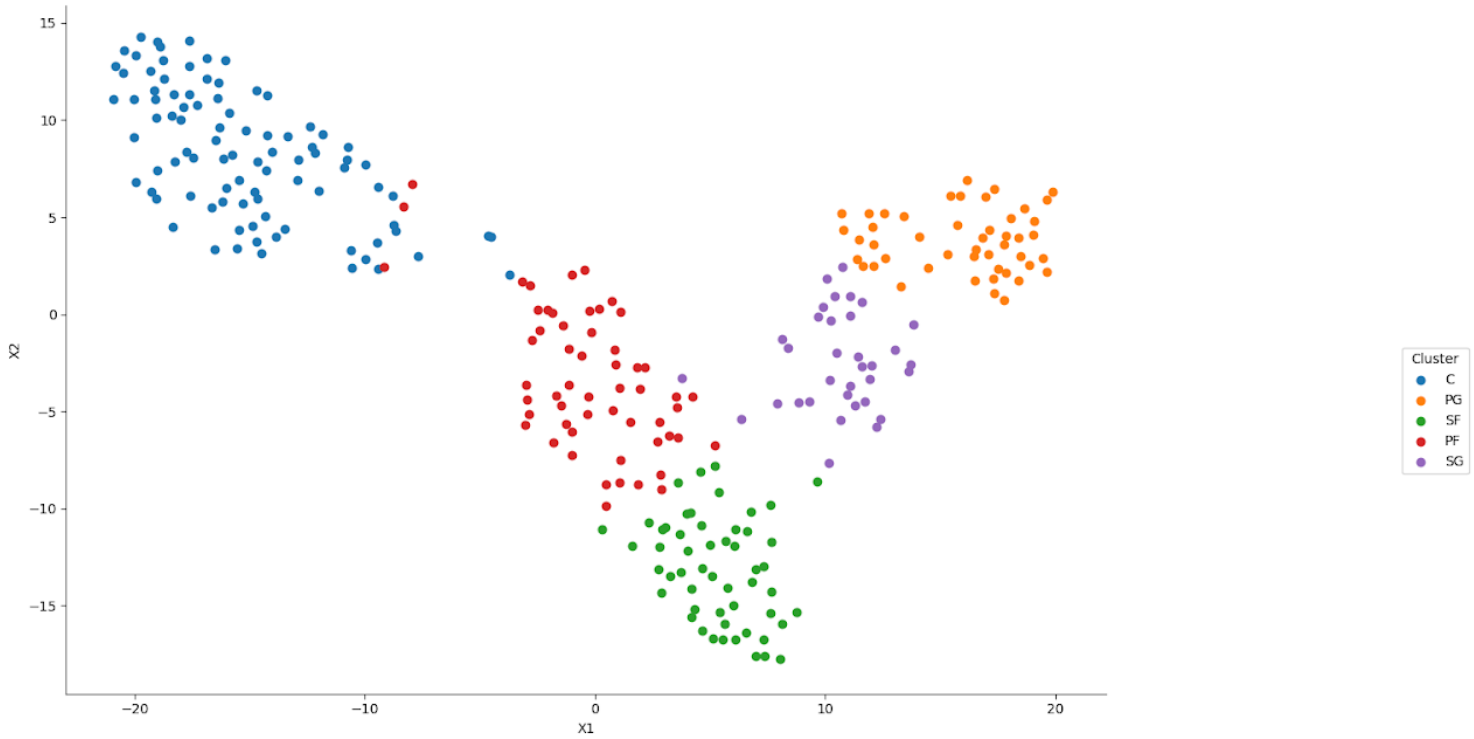
	A	B	C	D	E	F	G	H	I
1		player_name	player_id	FG_ranking	ScoAvg_ranking	sco_ranking	off_ranking	ALL_RANK_SCORE	ALL_RANK
2	109092	deandre jordan	201599	1	116	53	36	755	1
3	36622	tyson chandler	2199	2	126	54	62	854	2
4	59453	danny green	201980	183	91	10	166	872	3
5	35165	patrick patterson	202335	103	143	11	187	917	4
6	107109	matt barnes	2440	126	123	23	155	944	5
7	48426	chris copeland	203142	269	168	36	56	965	6
8	20013	kyle korver	2594	57	81	2	269	965	6
9	17336	draymond green	203110	157	100	28	150	987	8
10	8408	rudy gobert	203497	3	207	76	18	1024	9
11	28029	ryan anderson	201583	219	60	42	126	1077	10
12	104258	anthony morrow	201627	147	125	24	189	1079	11
13	89251	jonas jerebko	201973	97	224	35	141	1094	12
14	124368	mirza teletovic	203141	254	144	18	175	1103	13
15	115917	wesley matthews	202083	132	25	47	161	1110	14
16	50034	cj miles	101139	256	86	27	182	1158	15
17	41466	channing frye	101112	246	157	3	243	1162	16
18	89483	kyle singler	202713	243	184	13	204	1169	17
19	71063	robert covington	203496	241	85	17	228	1180	18
20	127097	bojan bogdanovic	202711	197	175	49	110	1192	19
21	29594	luke babbitt	202337	93	230	4	280	1203	20

II. Defender ranking result

1	player_name	player_id	FGM_rank	dribbles_rank	touch_time_rank	def_distance_rank	consensus_ranking
2	Williams Elliot	202343	11	16	6	5	1
3	Barron Earl	2853	16	1	1	16	2
4	Robinson Glenn	203922	22	119	183	41	3
5	Salmons John	2422	26	207	208	19	4
6	Dinwiddie Spencer	203915	97	54	64	29	5
7	Kuzmic Ognjen	203136	17	59	42	179	6
8	Calderon Jose	101181	31	76	57	154	7
9	Gordon Drew	204079	68	36	10	143	8
10	Williams Mo	2590	117	57	55	54	9
11	Bargnani Andrea	200745	13	338	290	67	10
12	Kirk Alex	203945	2	453	461	2	11
13	Christopher Patrick	203565	9	454	460	3	12
14	Whittington Shayne	203963	37	261	192	116	13
15	Douglas-Roberts Chris	201604	10	458	455	14	14
16	Price Ronnie	101179	127	56	53	94	15
17	Payton Elfrid	203901	139	63	71	66	16
18	Williams Reggie	202130	28	430	451	4	17
19	Shved Alexey	203144	44	198	304	111	18
20	Martin Kevin	2755	98	219	219	44	19
21	Wilcox CJ	203912	15	469	470	21	20

c. 利用 cluster 將 Data 裡面的球員分群：我們定義了以下幾個 feature 作為判斷一個球員在球場上位置的資訊，分別有，平均投籃距離、平均投籃次數、總得分、命中率、運球次數、持球時間、兩分球與三分球比例，總共八個 feature，而我們這邊利用 **Silhouette Coefficient** 這個 metric 來判斷 cluster 的好壞的結果，其數學式為$s = \frac{b-a}{max(a,b)}$，a 代表當下 sample 點與其 cluster 內的所有點的平均距離，b 則代表與 sample 點距離最近 cluster 內所有點與 sample 點的平均距離，因此可以從這個 metric 看出當 s 越趨近於正負 1 時，代表 cluster 分的越開，cluster 效過越好；我們發現當使用八個 feature 時，算出來的 s 值大約只有 0.2 左右，代表 overlap 的點數高，因此我們重新對 feature 做調整，當刪除到只剩下兩分與三分球比率及投籃距離的 feature 時，s 可以上升到將近 0.5，但實際看 cluster 與 true label 的比對，發現差異很大，代表雖然我們可以把 cluster 分的乾淨，但因為使用的 feature 太少，cluster 出來的結果與實際情況不符，而在試過多個結果，決定後來採用六個 feature，也就是從原本的八個 feature 中刪除總得分與運球次數的

feature，score 大約也有 0.4 左右，但預測值與實際 label 有較多相同，下圖為 cluster 後的結果並把結果與 true label 做成表格做比對



		player_id	player_name	Ground_truth	predict_label
2	0	708	kevin garnett	PowerForward,SmallForward,Center	C
3	1	977	kobe bryant	SmallForward,ShootingGuard	PG
4	2	1495	tim duncan	Center,PowerForward	C
5	3	1713	vince carter	SmallForward,ShootingGuard	SF
6	4	1717	dirk nowtizski	Center,PowerForward	PF
7	5	1718	paul pierce	ShootingGuard,SmallForward	PF
8	6	1889	andre miller	PointGuard,ShootingGuard	PG
9	7	1890	shawn marion	PowerForward,SmallForward	C
10	8	1891	jason terry	PointGuard,ShootingGuard	SF
11	9	1938	manu ginobili	ShootingGuard	SG
12	10	2034	mike miller	ShootingGuard,SmallForward	SF
13	11	2037	jamal crawford	ShootingGuard,PointGuard	SG
14	12	2045	hedo turkoglu	PowerForward,SmallForward	SF
15	13	2199	tyson chandler	Center,PowerForward	C
16	14	2200	pau gasol	Center,PowerForward	C
17	15	2207	joe johnson	ShootingGuard,SmallForward	SG
18	16	2210	richard jefferson	SmallForward	SF
19	17	2216	zach randolph	PowerForward,Center	C
20	18	2225	tony parker	PointGuard	PG
21	19	2365	chris andersen	Center,PowerForward	C
22	20	2403	nene hilario	Center,PowerForward	C

E. Conclusions

a. 基於進攻/防守的排名，我們可以再次審視某些球員的進攻/防守能力。我們平場會看的數據很難顯示出他在球場上進攻的動向是如何，我們有了每次進攻的時間以及防守距離等參數，可以將球員的強度評估考慮得更為周全。

b. 基於這些數據，我們做了球員的 clustering，有些球員可能因為身高/體重等等外部因素而分配打某個位子，但可能不適合這個人而導致他的成績低落，對球隊的勝利也有額外的影響。如果可以從他進攻/防守的這些數據來看，且屏除身高/體重這些外部因素，或許可以找到某些人在球場上更適合的位置。

c. 在進攻部分，我們引用了進攻效率當作參數，但我們發現看得分與持球時間的比率或許對某些位置的球員不太公平，例如控球後衛必須長時間持球組織進攻，或許平均的持球時間會比其他人還要長，造成這項排名比較後面；我們也引用了得分難度當作參數，但得分難度跟防守者的距離不一定是成反比，反倒是防守者的距離大到一定程度之後投籃難度就會急遽下滑，或許我們可以用更複雜的函數來表達這兩個參數。

d. Silhouette Coefficient 這個 metric 來調整我們的 feature，使的預測出來的 label 與 ground truth label 較為相近，但仍舊有些球員預測出來的位置與現實不符，而透過 metric 評斷的結果，我們可以根據現有的 data 將所有球員分成五群，而且 Silhouette Coefficient 的分數也不差，代表我們確實是可以將這些球員成功定義於五個不同的位置，因此我們想做的事情即是將這樣的數據提供給球團當作調整球員的依據，如果一位球員並不是真正熟悉現在打球的位置，或許可以透過這樣的分析資料幫助他們找到真正屬於自己的位置。

F. Future work

a. 由於在球場上的某些球員並不是只打唯一一個位置而已，因此我們必須考量到某些球員可以同時站兩個以上的位置的可能性，也就是所謂的「Multi-label」，所以我們可以加上 soft cluster，這樣就可以將球員同時標記在兩個以上的位置，並且更進一步去分析這些球員彼此之間的搭配。

b. 第二點要考慮到的是，在進攻/防守的排行上，可以觀察到的是我們所做出來的排行並不是和以往大家所熟悉的排行相似，位居前幾名的並不是大家所熟悉的明星球員，所以在我們的 feature 之中，也可以加入以往傳統的數據，ex: 抄截、火鍋、籃板...等等，來增強我們的排名的可信度。

c. 承第二點，我們可能需要一些數據來當作評斷我們 feature 好壞的標準，所以我們可以就進攻以及防守方面分別找當年的最佳防守球員或是最佳防守隊伍來當作我們的 metric，並研究這些球員在我們取的 feature 上是否有特別突出的地方，再對我們的 feature 做修正，應該會有比較好的結果。

G. Teamwork Assignment (if applicable)

A071547 鄭澄遠：數據處理、演算法、報告、PPT

A061502 陳文揚：數據處理、演算法、報告、PPT

A061618 黃有德：Spotlight Video、演算法、報告、PPT