

나만의 데이터 분석 End - to End 프로젝트 기획서

- 수비수의 contest 정도에 따라 슛 정확도에 유의미한 차이가 발생하는가 -

데이터반 남기동

1. 주제 의식 및 개요

개인적으로 즐겨보는 스포츠인 미국의 프로 농구(NBA)에서 수비수의 contest가 공격자의 슛 정확도에 유의미한 영향을 미치는지 분석해보고자 한다. Contest는 수비수가 슛을 쏘는 공격자에게 다가 압박을 하고 저지하는 것을 의미한다. 하지만 contested shot과 blocked shot은 차이가 있다. 수비수에 의해 저지가 된 것은 블락되었다(blocked)라고 표현하지만 수비수의 방해를 받았지만 슛을 던진 경우 contested shot이라고 표현한다. 이 때 수비수가 공격수의 공을 직접적으로 저지하지 못할 거리라는 것을 인지하고 있음에도 불구하고 공격자에게 압박을 가하기 위해 달려드는 수비를 하는 경우가 있다. 이러한 수비는 슛을 쏘는 공격자에게 심리적인 압박을 하기 위한 의도이지만 만약 공격자가 슛을 쏘는 것이 아니라 슛 페이크를 준 것이라면 수비가 뚫릴 수 있는 위험이 따른다.



이와 관련해서 개인적으로 농구를 직접 하는 것을 즐기기 때문에 플레이 도중 들었던 의문으로, 만약 심리적 압박을 목적으로 한 수비의 결과가 상대 공격수의 슛 정확도에 유의미한 차이를 이

끌어내지 않는다면, 무리해서 달려드는 수비를 할 필요가 없지 않을까 의문이 들었다. 물론 아마 추어의 경우 심리적 압박을 느끼는 것이 분명하지만 과연 세계 정상급 실력의 NBA 선수들은 과연 수비수의 contest에 유의미한 반응을 보일까 하는 의문이다.

2. 데이터 수집 방법(데이터 출처 포함)

<https://www.nba.com/stats/players/shots-closest-defender>

위의 링크에서 NBA에서 공식적으로 제공하는 선수들의 데이터를 볼 수 있다. 그 중에서도 Shot Dashboard 중 Closest Defender에 들어가면 수비수와의 거리에 따른 슛 시도와 슛 성공 횟수, 슛 성공률 등을 확인할 수 있다. 시즌은 2013-14부터 2022-23까지 10년치 정보를 조회할 수 있으며 season의 타입으로 Regular season과 playoff의 데이터를 조회할 수 있다. 특히나 'Closest Defender Distance Range' 옵션에서 네 가지 옵션을 선택할 수 있다.

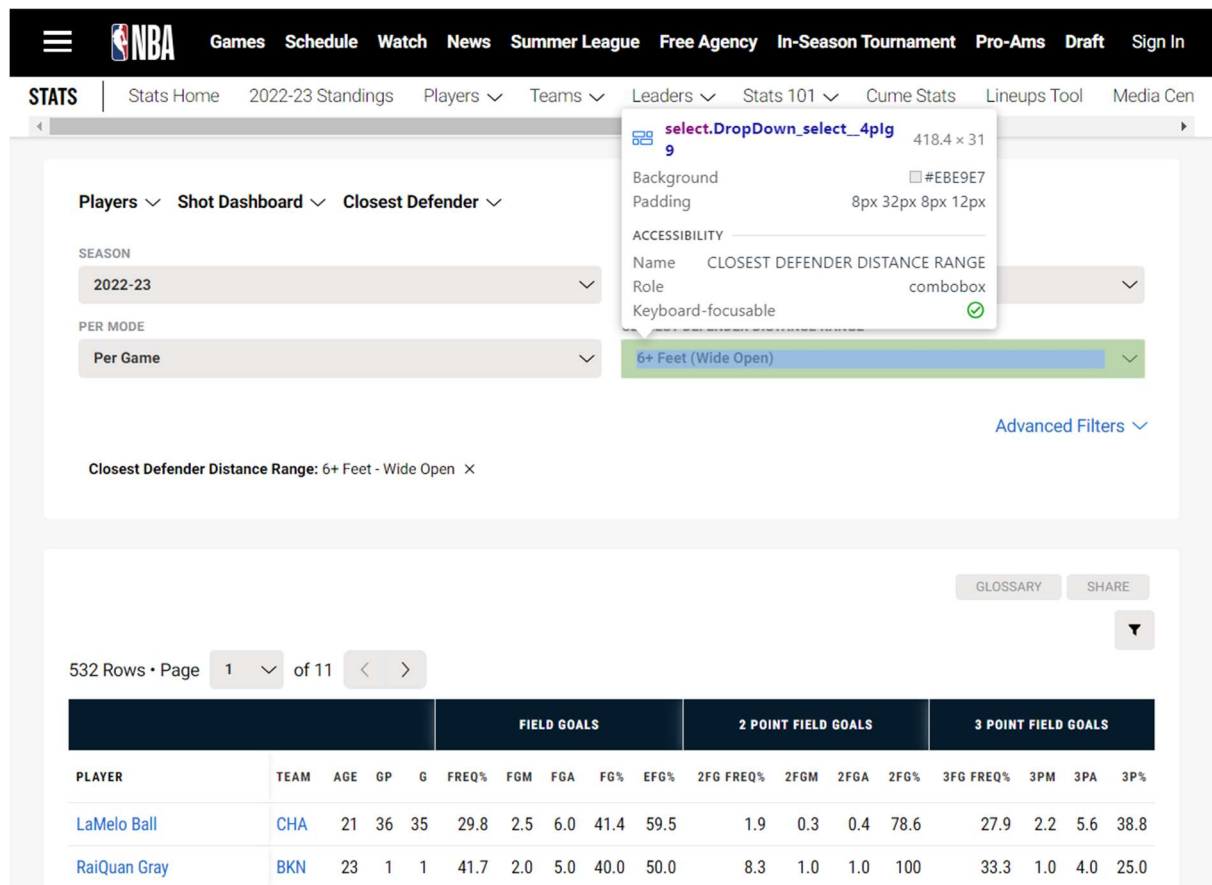
1. 0 – 2 Feet (Very Tight)

2. 2 – 4 Feet (Tight)

3. 4 – 6 Feet (Open)

4. 6+ Feet (Wide Open)

데이터의 수집은 데이터 크롤링을 통해 수행할 수 있다.



The screenshot shows the NBA Stats website interface. The top navigation bar includes links for Games, Schedule, Watch, News, Summer League, Free Agency, In-Season Tournament, Pro-Ams, Draft, and Sign In. The main content area is titled 'Stats' and includes a dropdown menu for 'Closest Defender' set to '6+ Feet (Wide Open)'. Below this, there are filters for 'SEASON' (2022-23) and 'PER MODE' (Per Game). The table displays statistics for 532 rows, with columns for Player, Team, Age, GP, G, FREQ%, FGM, FGA, FG%, EFG%, 2FG FREQ%, 2FGM, 2FGA, 2FG%, 3FG FREQ%, 3PM, 3PA, and 3P%. The first two rows shown are LaMelo Ball (CHA) and RaiQuan Gray (BKN).

PLAYER	TEAM	AGE	GP	G	FIELD GOALS					2 POINT FIELD GOALS				3 POINT FIELD GOALS			
					FREQ%	FGM	FGA	FG%	EFG%	2FG FREQ%	2FGM	2FGA	2FG%	3FG FREQ%	3PM	3PA	3P%
LaMelo Ball	CHA	21	36	35	29.8	2.5	6.0	41.4	59.5	1.9	0.3	0.4	78.6	27.9	2.2	5.6	38.8
RaiQuan Gray	BKN	23	1	1	41.7	2.0	5.0	40.0	50.0	8.3	1.0	1.0	100	33.3	1.0	4.0	25.0

각 세부 항목당 400~500개 행의 데이터가 있기 때문에 옵션을 자동적으로 바꾸고 테이블의 정보를 크롤링하는 코드를 작성하여 수집하면 될 것이다.

3. 예상되는 데이터 전처리 방법

NBA 공식 사이트에서 제공하는 만큼, 그리고 많은 열로 이루어져 있지 않은 데이터이기 때문에 결측치가 있는 경우는 거의 없을 것으로 예상된다. 그러나 앞서 설계한 프로젝트의 목표인 '수비수의 contest가 NBA에 선수들에게 유의미한 압박을 줄 수 있는가'를 분석하는데 있어서 데이터 전처리에 포함해야 할 중요한 부분이 있다. 그것은 슛 시도 자체가 적은 선수들의 데이터는 배제해야 한다는 것이다. 사실 정규 시즌에 경기를 뛴 선수의 데이터가 행으로 따졌을 때 500행 가량 되지만 실질적으로 contested shot을 자주 던지는 선수는 훨씬 적다. 그러나 경기 수가 적고 시도 횟수도 적은 선수들의 데이터를 포함하면 슛 성공률을 따지는 것에 있어서 원하지 않는 결과를 도출할 수 있다. 따라서 출전 경기 수가 10회 이하이거나 contested shot의 정규 시즌 시도 횟수가 10회도 되지 않는 선수들의 데이터들은 아예 제외하는 것이 좋을 것이다.

수집해야 할 핵심 데이터, 즉 실질적인 데이터는 동일 선수를 기준으로 수비 거리(0 ~ 6+ feet)에 따른 슛 정확도의 차이이다. 즉 데이터 자체에서 바로 가져올 수 없는 '차이'를 계산해야 하기 때문에 기준을 세우고 편차를 구하여 새롭게 데이터 테이블을 만들어야 한다. 이를 위해서 각 거리에 따른 선수들의 데이터를 수집한 다음 6+ feet(wide open)을 기준으로 하여 거리가 감소함에 따라 슛 정확도가 얼마나 줄어드는지 편차를 계산하여 데이터를 저장한다.

4. 데이터 분석 방법

동일 선수를 기준으로 수비 거리(0 ~ 6+ feet)에 따른 슛 정확도의 차이를 구하여 수비수의 contest가 유의미한 압박을 줄 수 있는지 판단하기 위해서 우선 수비수의 contest 정도(수비수와 거리)를 독립 변수로, 그에 따른 슛 정확도의 차이를 종속 변수로 하는 단순 선형 회귀(Simple Linear Regression) 모델을 설계하고 학습하여 수비수의 contest 정도가 공격자의 슛 정확도에 유의미한 영향을 미치는지 모델 결과를 해석하는 과정에서 판단하고자 한다. R-squared 값이 1과 가까울수록 회귀식이 원래의 자료를 잘 설명해준다는 사실에 근거하여 판단해볼 수 있을 것이다. 회귀 모델 실습에서 하였던 데이터 전처리를 수행하여 R-squared 값을 최대한 상승시켜 보고 그 때 과연 회귀식이, 모델이 유의미한 설명력을 가지고 있는지 판단해볼 수 있을 것이다.

5. 기대 효과 및 예상되는 결과

2022-23 시즌뿐만 아니라 10년치 데이터를 모두 전처리 하여 모델에 학습시켜 수비수의 contest

정도와 슛 정확도(성공률)의 관계, 나아가 유의미한 슛 정확도의 차이가 발생하는지 계산해볼 수 있다. 만약 이렇게 학습시킨 선형 회귀 모델의 R-squared 값이 0.7을 넘는다면 수비수의 contest 정보가 그래도 슛 정확도가 감소하는 것에 확실히 영향을 미친다는 것으로 이해해볼 수 있다. 하지만 그 정도가 어느 정도인지도 중요하다. 만약 wide open에서 시도하는 슛과 타이트한 상태에서 시도하는 슛의 성공률이 5%p도 차이가 나지 않는다면 NBA의 정상급 shooter를 상대하는 수비수 입장에서 무리한 contest를 시도하는 것이 비효율적이라 생각해볼 수 있을 것이다. 무리한 contest 시도 때문에 수비 구조가 흐트러져서 상대팀 다른 선수에게 좋은 기회가 넘어가는 경우도 많기 때문이다.

개인적으로는 단순 선형 회귀 모델을 설계하고 모델 결과를 해석한다면 contest 정도가 종속변수인 슛 성공률에 분명 유의미한 영향을 미칠 것이지만 그 영향의 정도가 5%p일 것이라 생각한다.

Wide open(6+ feet)에서 Kawhi Leonard는 56.1%, Jimmy butler는 58.8%, Kevin Durant는 54.3%의 슛 성공률을 갖는다. 반대로 0 – 2 Feet의 Very tight한 상태, 즉 수비수의 contest가 가장 심할 때 세 선수의 성공률은 Kawhi Leonard가 53.8%, Jimmy butler가 46.3%, Kevin Durant가 51.6%이다. 비록 Jimmy butler는 비교적 큰 차이가 있었지만 Kawhi Leonard와 Kevin Durant는 5%p 이하의 차이가 발생했다. 물론 이 선수들은 NBA에서도 최정상의 슈퍼스타이기 때문에 비교적 영향을 더 받았다고 할 수 있겠지만 평균적으로도 5%보다 큰 차이가 발생하지는 않을 것으로 예상된다.