

최종 칼럼

[야구 3조 칼럼.pdf](#)

1. 서론

마운드 위의 투수가 어떠한 공을 던질 것인가'에 관한 궁금증은, 수많은 야구팬들의 이목을 집중시키는 탐구 주제이다. 하지만 KBO 리그의 투수에 대한 평가는, 승패나 평균자책점과 같은 전통적 지표에 크게 의존하고 있는 현실이다. 이번 탐구에서는 이와 같은 문제점을 해결하고 투수 평가 지표 방식의 다양화를 위해, 'KBO 투구 데이터의 정량적 분석과 모델 구축'을 주제로 삼아 분석을 진행하였다. 본 탐구는, 기존 분석 방식의 한계로부터 벗어나기 위해, 단순 경기의 결과가 아닌 선발 투수들의 공을 한 구 단위로 수집하고, 이를 바탕으로 투구 스타일에 대한 분석을 제공할 수 있는 모델을 구축하고자 했다. 네이버 문자중계 크롤링을 통해 생성한 대규모 투구 로그를 기반으로 군집 분석, 투수 유형 분류, 결정구 및 투구 시퀀스 분석을 수행했으며, 초기 모델의 한계를 보완하기 위해 리그 전체 투구 데이터로 분석 범위를 확장했다. 본 연구는 투수 평가 지표의 다양화와 함께, KBO 투구 데이터를 활용한 정량적 분석 방법론의 가능성을 제시하는 데 목적이 있다.

2. 데이터 수집방법

본 분석은 2024년과 2025년 시즌 KBO 리그에서 연속으로 규정 이닝을 충족한 국내 선발 투수 11인을 대상으로 수행되었다. 데이터 수집을 위해 네이버 스포츠의 선발 등판 경기 로그를 기초 자료로 활용하였으며, 네이버 문자 중계 시스템을 크롤링하여 개별 투구 단위의 상세 정보를 확보하였다. 수집된 주요 특성(Feature)은 투구별 상대 타자, 누적 투구수, 구종, 구속, 볼카운트, 투구 결과 및 타격 결과 등을 포함하며, 표본의 신뢰성을 확보하기 위해 투수 1인당 약 4,500구에서 5,500구에 달하는 데이터를 구축함으로써 심층적인 투구 패턴 분석을 위한 기초 토대를 마련하였다.

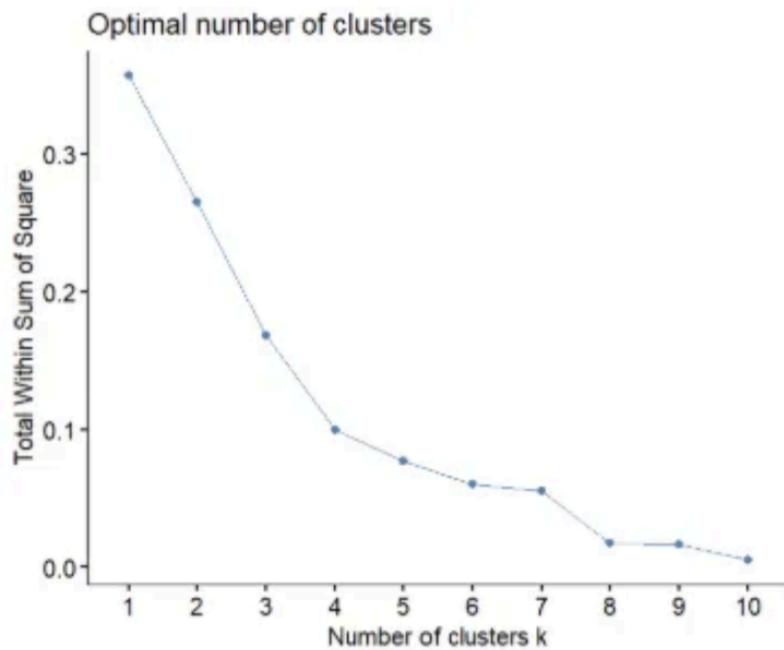
3. KBO 투수 구종 데이터 분석

3.1. 투구 스타일 군집 분석

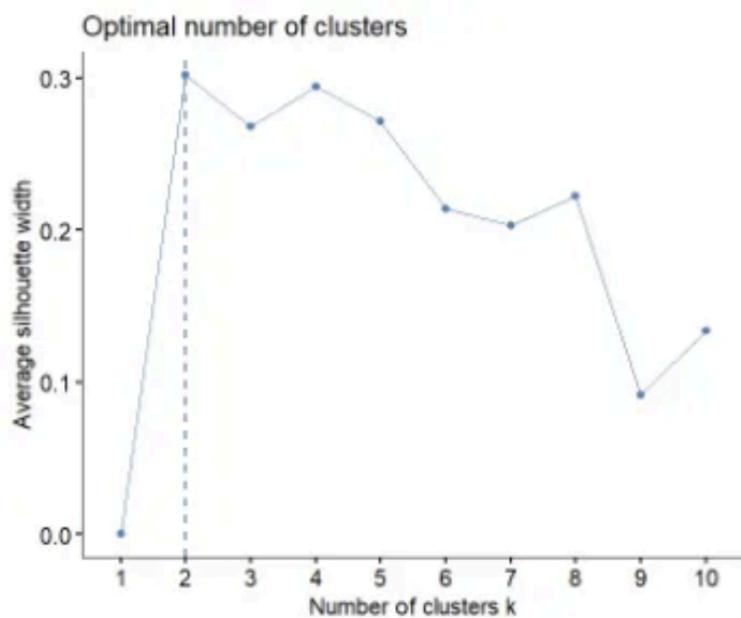
투수들의 투구 스타일을 객관적으로 분류하기 위해 군집 분석을 수행하였다.

먼저, 직구와 포심을 통틀어 직구, 투심, 커터, 슬라이더, 커브, 체인지업의 여섯 가지 범주로 구종을 분류하였다.

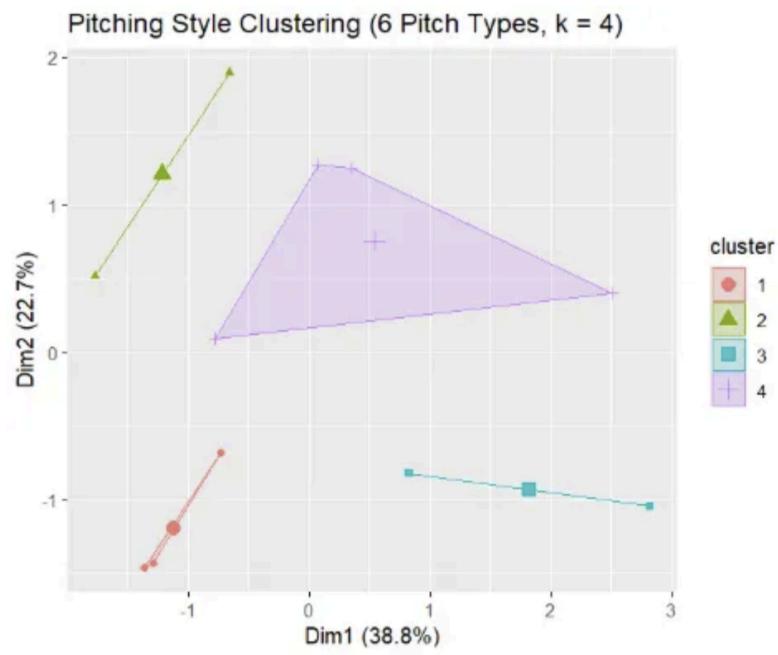
최적의 군집 개수 k 를 찾기 위해 엘보우 방법과 실루엣 방법을 수행하였다. 그 결과, 엘보우 포인트는 $k=4$, 실루엣은 $k=2$ 였다. $k=2$ 일 경우 분할의 정확성이 떨어질 우려가 있어 $k=4$ 로 결정하였다.



[엘보우 방법]



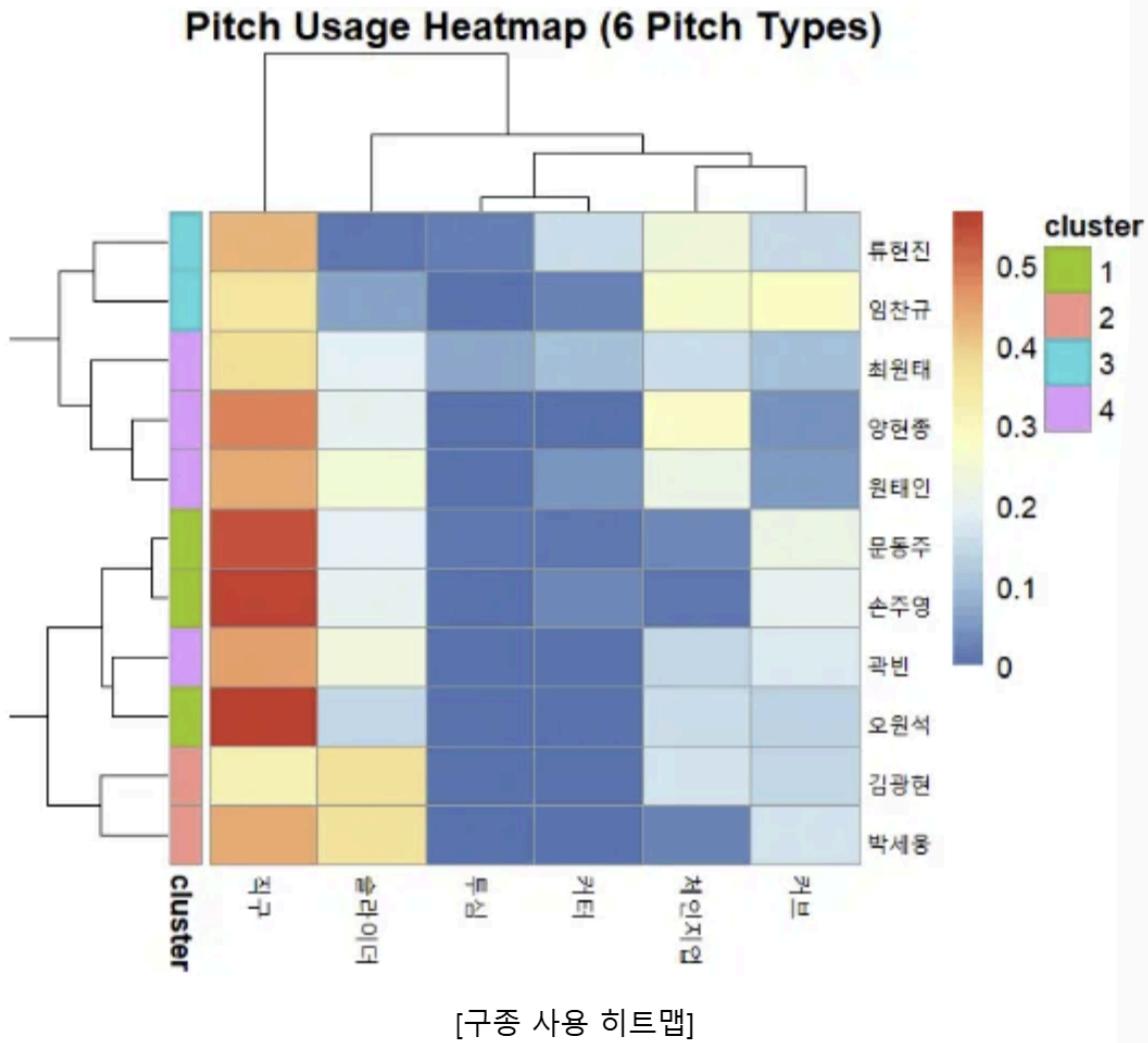
[실루엣 방법]



[클러스터링 시각화]

	cluster	슬라이더	직구	체인지업	커브	특심	커터
	<fct>	<db7>	<db7>	<db7>	<db7>	<db7>	<db7>
1	1	0.180	0.554	0.0635	0.186	0.00360	0.0135
2	2	0.369	0.379	0.0961	0.156	0	0
3	3	0.0364	0.392	0.254	0.215	0.0111	0.0913
4	4	0.216	0.435	0.198	0.0950	0.0184	0.0379

[클러스터별 구종 사용 비중 평균]



[구종 사용 히트맵]

군집 분석 결과는 다음과 같다.

Cluster 1: 직구 지배형 (문동주, 손주영, 오원석)

첫 번째 그룹은 히트맵에서 직구 영역이 진한 붉은색을 띠며, 슬라이더, 체인지업 등 다른 구종의 구사율이 현저히 낮다. 복잡한 수싸움보다는 패스트볼의 구위 하나로 타자를 찍어 누르는 스타일이다.

Cluster 2: 직구+슬라이더 병행형 (김광현, 박세웅)

한국 야구의 전통적인 에이스 유형이다. 이들은 강력한 직구를 바탕으로 하되, 슬라이더를 제2의 직구처럼 활용한다. 데이터를 보면 직구와 슬라이더의 비중이 이상적인 균형을 이루고 있다. 직구로 카운트를 잡고, 날카롭게 꺾이는 슬라이더를 결정구로 삼는다. 다양한 구종을 섞기보다는 확실한 구종들로 경기를 풀어가는 클래식의 전형이다.

Cluster 3: 체인지업 활용 완급형 (류현진, 임찬규)

이 그룹의 핵심 무기는 체인지업이다. 직구 구속이 Cluster 1만큼 빠르지 않더라도, 체인지업을 이용해 타자의 타이밍을 무너뜨린다. 힘보다는 타이밍 싸움으로 타자를 압도한다.

Cluster 4: 슬라이더/커터 특화형 (곽빈, 양현종, 원태인, 최원태)

이 그룹은 슬라이더와 커터의 비중이 매우 높다. 홈플레이트 근처에서 지저분하게 휘어지는 무브먼트를 적극적으로 활용한다. 직구 구위도 뛰어나지만 변화구를 섞어 타자의 배트 중심을 피해 가는 능력이 탁월하다. 다양한 궤적을 그리는 공으로 범타를 유도하는 데 특화된 투수들이다.

3.2. 투수 유형별 분석

본 연구에서는 자료 수집을 통해 축적한 11명의 토종 국내 선발 투수의 2024-2025 시즌간의 투구 데이터를 바탕으로, 각 선수들이 어떤 유형에 속하는지 분류해보았다. 선수 분류를 위한 선수 스타일의 정량화를 위해, Power, Arsenal, Tempo, Impact, Command, Stamina, Stress Relief의 7가지 지표를 정의하였다. 이를 바탕으로 투구 내용에서의 핵심 지표를 근거로 하여 군집 분석을 수행했고, 그 결과 분석 대상의 11명의 투수를 3가지 유형으로 분류할 수 있었다.

첫번째 분류 유형은 강력한 구위형이다. 이는 구속과 위력으로 타자를 압도하는 선수를 의미한다.

이 유형에 해당하는 투수들은 power 지표와 impact 지표에서 두드러진 값을 보였다. 구위형 투수들은 강력한 구위를 바탕으로, 전체 투구 수 대비 헛스윙 비율을 높게 가져가고 2스트라이크 이후 결정적 상황에서 타자의 반응을 이끌어내는 능력이 뛰어나다는 공통적인 특징을 갖는다. 해당 유형에 속하는 선수로는 문동주, 김광현, 박세웅 등이 있으며, 이들은 빠른 구속과 묵직한 구질 기반의 투구로, 패턴을 단순하게 가져가더라도 높은 위력을 보이며 타자를 상대한다. 다만 Arsenal과 Stress Relief 면에서 대조적으로 낮은 값을 보이는 경우가 존재하여, 경기 후반부 컨디션 저하에 대한 우려를 함께 내포하고 있다는 것을 확인할 수 있다.

다음으로는 테크니컬 팔색조형이다. 이는 구속에서 비롯된 구위가 아닌, 투구 조합과 속도에서 강점을 보이는 선수 유형이다. Arsenal과 Tempo, Deception에서 높은 수치를 기록한 선수들이 이에 속하며, 대표적인 선수로는 임찬규, 양현종, 오원석 등이 있다. 이들은 같은 구속대의 투수들에 비해 다양한 구종 분포와 구속 변화를 활용하여 타석에 선 상대 선수의 타이밍을 빼앗는다. 랜덤 포레스트 모델을 통해 계산한 구속 변화 중요도에서 특히 높은 수치를 나타냈으며, 이는 투구 상황에서의 완급 조절 능력이 뛰어나다는 것을 의미한다.

마지막 투수 유형은 안정적인 밸런스형이다. 이들은 전반적인 완성도가 높은 선발 투수 유형이다.

안정적 밸런스형 투수들은 위 두 유형의 선수들과 달리, 특정 지표에서 뛰어난 모습을 보이는 것이 아니라 대부분의 지표에서 고른 모습을 보인다. Power, Command, Impact 등의 수치가 모두 평균 이상을 기록하며, 경기 흐름에 따라 크게 흔들리지 않는 투구패턴이 특징

이다. 이 유형에 속하는 선수로는 류현진, 원태인, 최원태, 꽈빈 등이 있으며, 이들은 초반과 후반의 헛스윙 비율에 큰 차이가 없다는 점에서, Stamina 지표에서 안정적인 값을 보인다는 공통점을 갖는다. 이는 곧, 전반적인 경기 운영과 투구 선택의 일관성이 반영된 결과로 해석 할 수 있다.

투수 유형별 분석은, 단순히 투수를 평가 및 서열화하기 위한 것이 아니다. 선발투수라는 같은 포지션을 갖는 선수라도, 선수 개개인이 각자 가지고 있는 강점과 경기를 풀어나가는 방식에는 분명한 차이가 존재한다. 이번 유형별 분석은, 투구 스타일에 대한 직관적인 이해와 유형 분석을 넘어, 실제 데이터를 바탕으로 선수들의 유형을 분석할 수 있다는 점에서 의의를 갖는다.

3.3. 투수 결정구 분석

XGBoost는 여러 개의 약한 예측 모델을 결합해 강력한 성능을 내는 양상블 학습 알고리즘이다. 이전 모델의 오차를 다음 모델이 점진적으로 수정해 나가는 방식을 기반으로 하며, 컴퓨터 자원을 효율적으로 사용하여 계산 속도가 매우 빠르다. 특히 모델이 복잡해져 발생하는 오류를 효과적으로 방지하여, 정형 데이터 분석에서 가장 높은 예측 성능을 보여주는 모델 중 하나이다.

본 연구는 XGBoost 모델을 활용하여 투수가 볼카운트 변화에 따라 느끼는 심리적 압박감과 아웃카운트 확보를 위한 전략 변화를 정량적으로 분석하고자 했다. 이를 위해 경기 데이터에서 2스트라이크 이후의 볼카운트별 피장타율을 산출했으며, 특정 카운트에서 장타 허용 빈도가 높아지는 지표를 근거로 각 상황이 투수에게 부여하는 위험도와 압박감 가중치를 설정했다. 결과적으로 풀카운트에 근접할수록 실점 억제를 위한 투구 선택이 장타율 변동에 미치는 상관관계를 도출했으며, 이를 통해 볼카운트별 전략적 중요도를 수치화된 가중치로 체계화했다.

이어 XGBoost 알고리즘을 활용하여 투수별 결정구의 '기대 성공 확률'을 정밀하게 산출했다. 해당 모델은 투수 개별 성향과 구종, 구속 간의 비선형적 상관관계를 학습하여 특정 상황에서의 통계적 위력을 정량화하도록 설계했다. 이때 범타나 삼진 등 투수에게 유리한 결과는 '성공(1)'으로, 안타·홈런·볼넷 등 부정적인 결과는 '실패(0)'로 정의하여 각 구종의 효과성을 학습시켰다.

pitcher	슬라이더	직구	체인지업	커브	커터	투심	포크
곽빈	18.14	34.36	19.38	28.12	0	0	0
김광현	41.6	22.18	9.99	25.55	0	0	0.69
류현진	1.89	45.91	23.5	10.93	15.32	1.83	0.61
문동주	15.16	42.91	3.46	17.2	0.18	0.27	20.83
박세웅	28.18	32.29	1.41	19.77	0	0	18.35
손주영	14.46	47.21	0.63	21.37	2.16	0	14.18
양현종	24.9	36.42	32.96	5.72	0	0	0
오원석	12.54	51.38	12.13	22.15	0	0.08	1.71
원태인	24.97	44.26	21.63	3.27	5.67	0.2	0
임찬규	5.12	28.46	35.33	26.85	3.58	0	0.66
최원태	10.73	38.48	20.77	21.9	5.06	3.05	0

: 단순 사용 비율이 가장 높은 구종

[투수별 2스트라이크 이후 구종 선택 비율]

모델링의 정밀도를 제고하기 위해 투수가 결정구 상황에서 사용하는 구종들을 투구 빈도에 따라 1, 2, 3순위로 체계화했다. 이는 투수가 실제 경기에서 가장 신뢰하는 주무기와 보조 구종의 실제 효율성을 비교하기 위함이었다. 분석 과정에서는 개별 투수의 구종별 평균 구속을 모델에 대입하여 도출된 예측 확률을 바탕으로, 투수 간 동일 순위 구종들의 성공률을 상호 비교하는 방식을 채택했다. 이를 통해 각 투수가 보유한 결정구의 등급별 가치를 객관적으로 평가했으며, 리그 내 타 투수들의 동일 순위 주무기 대비 경쟁력을 진단하는 전략적 지표를 도출했다.

1순위				2순위				3순위			
투수	순위	구종	성공률	투수	순위	구종	성공률	투수	순위	구종	성공률
손주영	1순위	직구	0.6398	임찬규	2순위	직구	0.6584	곽빈	3순위	체인지업	0.6276
김광현	1순위	슬라이더	0.6218	박세웅	2순위	슬라이더	0.5819	손주영	3순위	슬라이더	0.5984
원태인	1순위	직구	0.5982	류현진	2순위	체인지업	0.5805	양현종	3순위	슬라이더	0.5879
류현진	1순위	직구	0.5944	손주영	2순위	커브	0.5799	오원석	3순위	슬라이더	0.5637
최원태	1순위	직구	0.5867	곽빈	2순위	커브	0.5793	임찬규	3순위	커브	0.5578
오원석	1순위	직구	0.5785	원태인	2순위	슬라이더	0.5736	박세웅	3순위	커브	0.5558
임찬규	1순위	체인지업	0.5715	양현종	2순위	체인지업	0.5649	최원태	3순위	체인지업	0.5509
양현종	1순위	직구	0.5683	김광현	2순위	커브	0.5621	김광현	3순위	직구	0.5368
박세웅	1순위	직구	0.5636	오원석	2순위	커브	0.5052	류현진	3순위	커터	0.5308
문동주	1순위	직구	0.5405	문동주	2순위	포크	0.4947	원태인	3순위	체인지업	0.5249
곽빈	1순위	직구	0.5132	최원태	2순위	커브	0.4287	문동주	3순위	커브	0.5022

[결정구 기대 성공 확률]

3.4. 투구 시퀀스 분석

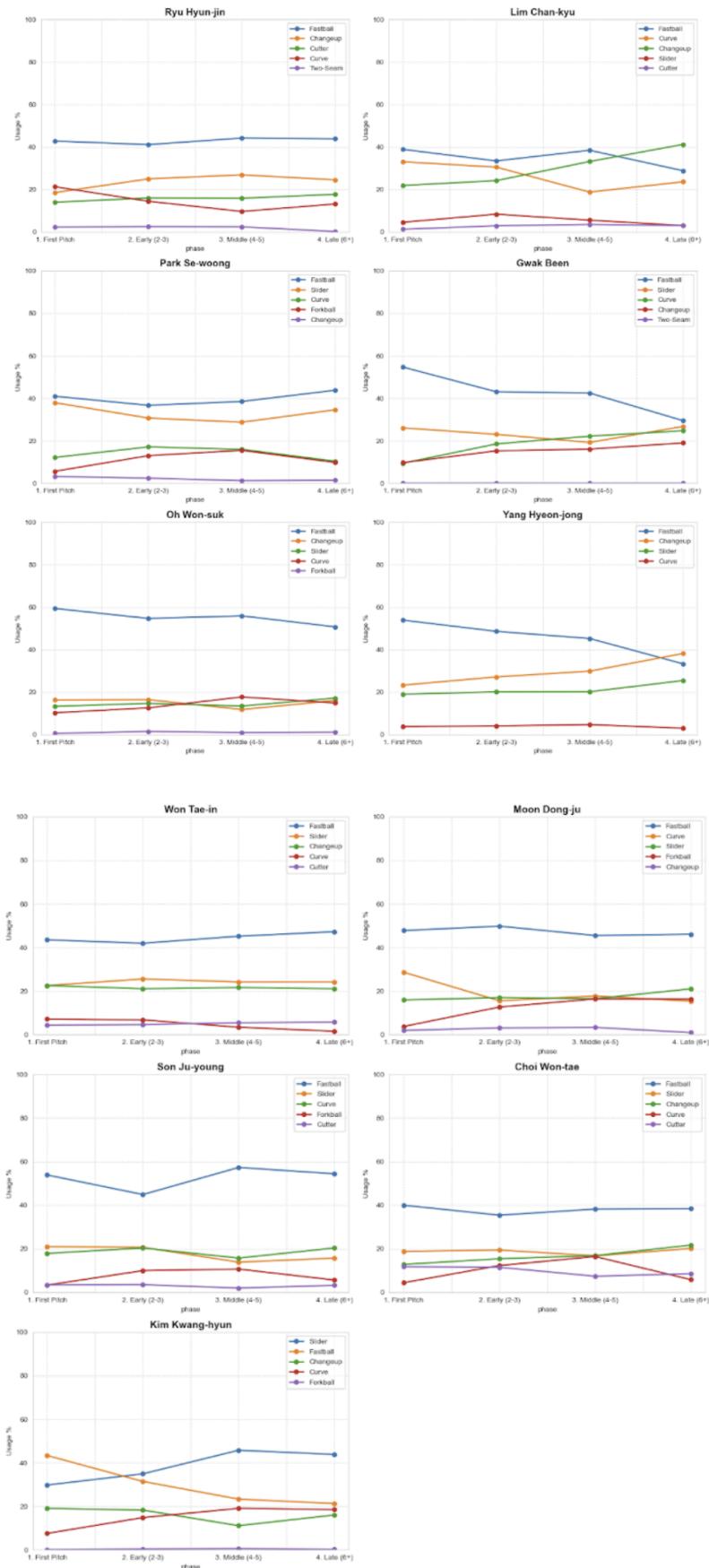
본 분석에서는 투수가 한 타석에서 투구 수에 따라 구종 선택 전략을 어떻게 바꾸는지를 알아 보고자 하였다. 1구에서 시작하여 6구 이상으로 길어지는 과정을 초반-중반-후반의 3단계로 나누어 구종 선택 전략을 추적하였다.

구종의 변화를 정량적으로 측정하기 위해 다음 네 가지 핵심 지표를 설계하였다.

1. 초구 직구 비율: 투수의 기선 제압 의지를 보여주는 지표이다. 수치가 높을수록 일단 힘으로 누르고 들어가는 공격적인 투수이다.
2. 결정구: 타자가 끈질기게 승부해 올 때 (6구 이상), 투수가 가장 많이 선택하는 최후의 보루가 무엇인지를 나타낸다.
3. 결정구 사용 비율: 그 결정구를 얼마나 확신을 가지고 던지는지를 보여준다. 비율이 높을수록 특정 구종에 대한 의존도가 높음을 의미한다.
4. 직구 비율 변화량: (후반 직구 비율) - (초구 직구 비율)로 계산한다. 이 값이 음수이면 승부가 길어질수록 도망가는 피칭을, 0 이상이면 끝까지 정면 승부를 하는 피칭을 의미한다.

Pitcher	First_Pitch_Fastball_Rate	Late_Count_Weapon	Late_Weapon_Usage_Rate	Fastball_Diff (Late - First)
Ryu Hyun-jin	42.8	Fastball	43.9	1.1
Lim Chan-kyu	38.9	Changeup	41.2	-10
Park Se-woong	41	Fastball	43.8	2.8
Gwak Been	54.7	Fastball	29.5	-25.2
Oh Won-suk	59.4	Fastball	50.7	-8.7
Yang Hyeon-jong	53.9	Changeup	38.2	-20.6
Won Tae-in	43.5	Fastball	47.3	3.7
Moon Dong-ju	47.8	Fastball	46.1	-1.7
Son Ju-young	53.9	Fastball	54.4	0.5
Choi Won-tae	40	Fastball	38.4	-1.6
Kim Kwang-hyun	43.4	Slider	43.9	-22.1

[투수별 스탯]



[투구 시퀀스 시각화]

투구 수가 많아짐에 따라 구종 선택에 큰 변화를 보인 투수들은 꽈빈, 김광현, 양현종이었다. 이들은 승부가 길어지면 직구를 과감히 버린다. 가장 극적인 투수는 꽈빈이다. 초구에는 54.7%라는 압도적인 비율로 직구를 뿐이며 힘으로 제압하려 들지만, 타자가 물러서지 않고 버티면 직구 비율을 29.5%까지 급격히 떨어뜨린다. 초반 대비 무려 25.2%P나 줄어든 수치다. 김광현과 양현종도 마찬가지다. 김광현은 장기전에서 직구를 줄이고 슬라이더를 결정구로 선택한다. 양현종 역시 직구 비중을 줄이며 체인지업으로 타자의 타이밍을 빼앗는다. 임찬규 또한 직구를 10%P 줄이고 체인지업의 비중을 높이며 위 선수들과 비슷한 성향을 보였다.

반면, 타자가 아무리 끈질기게 괴롭혀도 자신의 가장 강력한 무기를 절대 놓지 않는 투수들이 있다. 문동주, 최원태, 원태인, 손주영, 박세웅이다. 문동주와 최원태는 초구와 장기전의 직구 비율 차이가 거의 없다. 원태인은 장기전에서 오히려 직구 비율을 3.7%P 높이며 가장 공격적인 성향을 드러냈다. 손주영과 박세웅 역시 위기 상황에서 직구 비중을 높였다. 오원석은 직구 비중을 소폭 줄이긴 했으나 여전히 장기전에서 직구를 1순위로 선택하며 위 선수들과 비슷한 성향을 보였다.

그리고 이 두 그룹 어디에도 속하지 않는 투수가 있다. 바로 류현진이다. 초구와 장기전의 직구 비율 차이가 1.1%P로 거의 차이가 나지 않는다. 하지만 이는 직구만 던져서가 아니다. 그는 장기전에서도 직구, 체인지업, 커터 등 다양한 구종을 초구와 비슷한 비율로 밸런스를 맞춰 던진다. 특정 패턴에 얹매이지 않고 그때그때 타자의 허를 찌르는, 데이터로도 예측 불가능한 투수이다.

3.5. 주요 투수 예측 모델

본 분석은 투수별 투구 데이터에 XGBoost 알고리즘을 적용하여 다음 투구 구종을 예측하는 모델을 구축하는 것을 목적으로 한다. 분석의 효율성을 위해 개별 구종은 패스트볼, 브레이킹볼, 오프스피드볼의 세 가지 범주로 통합하였다. 분석 단위는 직전 투구, 타석, 당해 이닝, 당해 게임의 네 단계로 설정하여 다각적인 접근을 시도하였다. 모델 학습 과정에서는 shift 연산을 활용하여 직전 투구의 종류와 결과라는 시계열적 특성을 반영함으로써 데이터의 맥락적 의미를 강화하였다. 또한, 특정 구종에 데이터가 편향되는 불균형 문제를 해소하기 위해 compute_sample_weight를 적용하여 소수 클래스에 높은 가중치를 부여함으로써 모델이 희소한 투구 패턴에도 민감하게 반응하도록 설계하였다. 다만 초기 모델의 경우, 정밀도와 재현율의 조화평균인 F1-Score 기준 실질적인 예측력이 다소 낮게 도출됨에 따라 성능 최적화를 위한 추가 분석이 요구되었다.

직전투구 기반

Pre+Rec

pitcher	Acc	Pre	Rec	F1	max_depth	learning_rate	subsample	colsample_bytree	min_child_weight	reg_lambda
양현종	0.482252	0.469124	0.47391	0.470712	6	0.05	0.7	0.8	3	1
문동주	0.464497	0.479257	0.504492	0.455542	6	0.1	0.7	0.7	1	5
김광현	0.466447	0.453326	0.456786	0.435445	3	0.1	0.6	0.7	1	1
임찬규	0.440835	0.443548	0.440004	0.434607	4	0.05	0.8	0.7	3	10
곽빈	0.472956	0.432376	0.436834	0.431351	4	0.03	0.7	0.9	5	10
손주영	0.469626	0.436572	0.493068	0.429473	3	0.1	0.8	0.8	3	10
오원석	0.42464	0.436527	0.476012	0.420364	3	0.1	0.6	0.7	1	1
박세웅	0.423423	0.438553	0.463511	0.414706	3	0.1	0.6	0.7	1	10
원태인	0.428571	0.416744	0.415341	0.409618	3	0.1	0.8	0.8	3	10
류현진	0.43018	0.430641	0.454296	0.404829	6	0.1	0.7	0.7	1	5
최원태	0.415014	0.416895	0.422995	0.396058	6	0.1	0.7	0.7	1	5

[직전투구 기반 예측 결과]

타석 기반

Pre+Rec

pitcher	Acc	F1	Pre	Rec	subsample	max_depth	learning_rate	colsample_bytree
양현종	0.491803	0.48301	0.481162	0.493493	0.8	4	0.1	0.7
최원태	0.496823	0.462739	0.474214	0.489958	0.8	4	0.03	0.7
오원석	0.464871	0.454236	0.466792	0.503812	0.8	4	0.03	0.8
곽빈	0.475596	0.451523	0.451118	0.46863	0.7	4	0.03	0.7
김광현	0.478818	0.447113	0.453863	0.461183	0.7	5	0.1	0.8
손주영	0.481598	0.431352	0.439125	0.479861	0.8	4	0.03	0.7
박세웅	0.443439	0.427015	0.443324	0.471387	0.7	4	0.03	0.8
임찬규	0.431557	0.426959	0.434623	0.430817	0.7	5	0.05	0.8
문동주	0.418206	0.414174	0.422819	0.477485	0.7	4	0.03	0.8
원태인	0.423764	0.400089	0.406784	0.404727	0.8	6	0.03	0.7
류현진	0.432896	0.396907	0.410975	0.424749	0.8	4	0.1	0.8

[타석 기반 예측 결과]

당해이닝 기반

Pre+Rec

pitcher	Acc	F1	depth	lr	sub	col	lambda
양현종	0.512568	0.507859	4	0.1	0.6	0.7	1
김광현	0.495567	0.467136	4	0.05	0.7	0.6	1
곽빈	0.484677	0.46187	5	0.01	0.6	0.7	1
최원태	0.481576	0.452164	4	0.01	0.9	0.7	5
오원석	0.4637	0.451878	4	0.01	0.9	0.7	5
박세웅	0.451584	0.438614	5	0.01	0.6	0.7	10
원태인	0.443743	0.425339	5	0.05	0.8	0.9	5
임찬규	0.425287	0.423222	8	0.05	0.6	0.7	1
문동주	0.420844	0.417745	4	0.03	0.6	0.9	1
손주영	0.45531	0.415103	4	0.01	0.7	0.9	5
류현진	0.422805	0.389133	5	0.1	0.6	0.7	10

[당해이닝 기반 예측 결과]

당해게임 기반

Pre+Rec

pitcher	Acc	F1	Pre	Rec	subsample	reg_lambda	max_depth	learning_rate	colsample_bytree
양현종	0.535519	0.536217	0.535216	0.560361	0.7	10	4	0.03	0.7
김광현	0.495567	0.464618	0.471166	0.480978	0.7	1	4	0.03	0.7
곽빈	0.476731	0.464081	0.464393	0.500954	0.7	10	4	0.05	0.8
오원석	0.4637	0.447411	0.452413	0.486257	0.7	10	6	0.03	0.7
박세웅	0.460633	0.444991	0.449855	0.489271	0.8	10	4	0.1	0.8
문동주	0.44723	0.442503	0.460851	0.518051	0.8	10	4	0.05	0.8
최원태	0.468869	0.438428	0.454196	0.471358	0.7	1	4	0.1	0.7
손주영	0.488959	0.437032	0.442891	0.502438	0.8	10	6	0.03	0.7
임찬규	0.433647	0.431148	0.433829	0.434033	0.7	10	4	0.03	0.7
원태인	0.435331	0.405538	0.405135	0.406786	0.7	10	6	0.03	0.8
류현진	0.411705	0.401237	0.437046	0.461153	0.7	10	4	0.05	0.8

[당해게임 기반 예측 결과]

모델의 성능 개선을 위해 API 크롤링을 통해 2024-2025 KBO 리그의 406,479구에 달하는 방대한 투구 데이터를 추가로 확보하였다. 기존 변수 외에도 초기 전진 속도 성분, 플레이트 도달 시간, 수직 가속도 등 물리량 기반의 신규 특성을 수집하여 분석의 정밀도를 높였다. 이러한 물리적 특성을 바탕으로 구종 분류 모델을 재구축한 결과, 패스트볼, 브레이킹볼, 오프스피드볼 전 범주에서 우수한 정밀도와 재현율을 기록하며 96%의 높은 예측 정확도를 달성하였다. 특히 변화구와 오프스피드 계열에서도 안정적인 성능을 보여 모델의 신뢰성을 확보하였으며, XGBoost 학습 과정에서 검증 손실(Log Loss)이 반복 횟수에 따라 안정적으로 감소하는 우수한 학습 곡선을 확인하였다.

==== Classification Report (Validation) ===				
	precision	recall	f1-score	support
Fastball	0.98	0.98	0.98	43123
Breaking	0.93	0.96	0.94	23774
Offspeed	0.96	0.92	0.94	13968
accuracy			0.96	80865
macro avg	0.96	0.95	0.95	80865
weighted avg	0.96	0.96	0.96	80865

==== Classification Report (Validation) ===				
	precision	recall	f1-score	support
Fastball	0.98	0.98	0.98	43123
Breaking	0.93	0.96	0.94	23774
Offspeed	0.96	0.92	0.94	13968
accuracy			0.96	80865
macro avg	0.96	0.95	0.95	80865
weighted avg	0.96	0.96	0.96	80865

[XGBoost 기반 구종 분류 모델의 검증 결과]

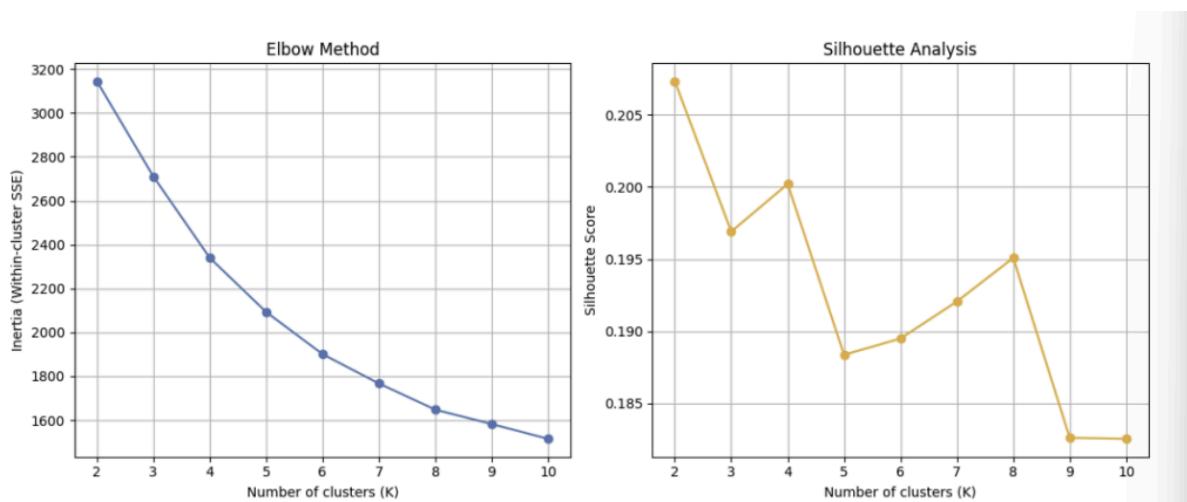
실질적인 차기 투구 예측을 위해 다양한 가중치 조합을 반복 실험하였으며, F1-Score를 최대로 끌어올리는 최적의 하이퍼파라미터를 탐색하였다. 최적화된 모델을 적용한 결과, 구종 별 예측률은 패스트볼 60%, 브레이킹볼 48%, 오프스피드볼 34%를 기록하였다. 파라미터 중요도 분석 결과에서는 직전 투구의 구종 그룹과 그 이전(두 번째 전) 투구의 구종 그룹이 가장 높은 중요도를 나타내었다. 이는 투수가 투구 시 공의 구속, 궤적, 릴리스 포인트 등 의 물리적 요소뿐만 아니라 이전 투구와의 연계성을 고려하여 구종을 조합한다는 전략적 특성을 시사한다.

==== Classification Report (Holdout) ====				
	precision	recall	f1-score	support
Fastball	0.62	0.58	0.60	43123
Breaking	0.43	0.55	0.48	23774
Offspeed	0.40	0.30	0.34	13968
accuracy			0.52	80865
macro avg	0.49	0.48	0.48	80865
weighted avg	0.53	0.52	0.52	80865

[최적 하이퍼파라미터를 적용한 XGBoost 모델의 최종 성능]

투구 데이터를 군집 분석하기 위해 데이터의 차원을 축소하기 위한 처리가 필요했고, 고차원 데이터의 효율적인 처리를 위해 주성분 분석(PCA)을 시행하였다. PCA는 데이터의 정보 손실을 최소화하면서 분산이 가장 큰 방향으로 새로운 축을 설정하여 변수 간 상관관계를 독립적인 성분으로 재구성하는 기법이다. 이를 통해 복잡한 데이터 구조를 단순화하고 핵심 특징을 보존하며 연산 효율성을 제고하였다. 분석 전에는 비율 데이터의 합 제약 조건을 해결하고 변수 간 독립성을 강화하기 위해 각 성분을 기하평균 대비 로그 비율로 변환하는 CLR(Centered Log-Ratio) 변환을 선행하였다. 최종적으로 PCA를 통해 도출된 주성분을 바탕으로 군집 분석을 수행하였으며, 응집도를 나타내는 inertia와 분리도를 나타내는 silhouette 점수를 종합적으로 고려하여 최적의 군집 수를 결정하였다.

하단의 시각 자료는 7개의 주성분을 바탕으로 투수를 군집분석한 결과이다. 각각의 주성분은 구종 비율, 속도, 성분, 위치, 변수등의 특성을 가지고 있다.



[군집 개수 결정]

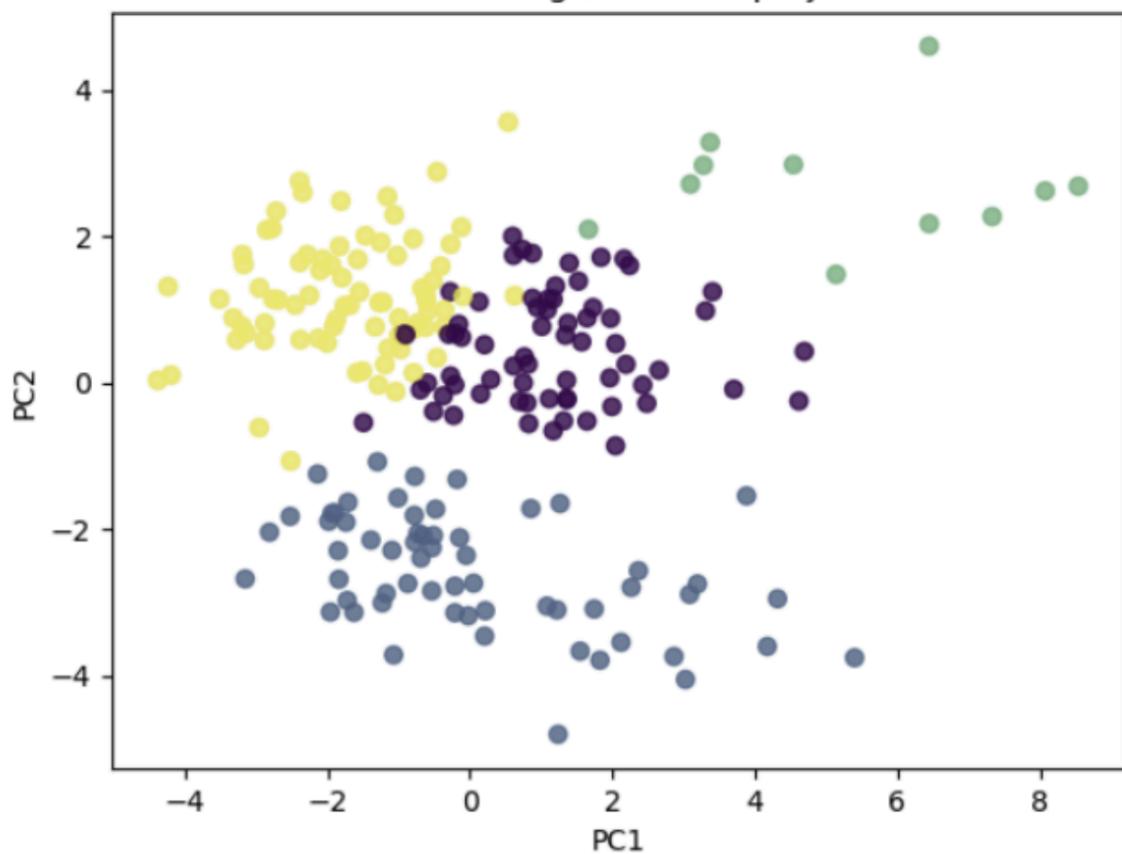
pitcherId	PC1	PC2	PC3	PC4	PC5	PC6	PC7	cluster
50030	3.704672	-0.08985	-0.83542	2.698763	-2.94006	0.815168	-1.76113	0
50106	-2.71544	2.342735	-1.96556	0.990573	0.703939	0.202142	-1.16497	3
50109	0.05381	-2.74106	-1.0045	-2.31285	0.664448	-0.06848	-0.96462	1
50354	-1.62459	-3.14009	-1.60456	-1.60243	-0.88959	0.39819	-0.1447	1
50393	1.203678	1.322219	-0.06931	-0.37522	-0.42695	-1.57533	0.397951	0
50441	0.878547	1.767839	-0.95061	0.604371	-0.69095	-0.5788	0.836715	0
50464	-0.14173	-2.11825	-0.98652	-2.68994	0.03103	0.50791	-0.28915	1
50556	0.131538	1.104241	1.617899	0.729894	-0.80046	1.84783	-0.4487	0
50662	-2.87657	0.811855	-0.98496	2.068991	-1.40262	0.221723	-0.23368	3
50812	0.858416	-1.72024	-1.19866	0.575678	-0.23638	-0.91658	1.475395	1
50859	-0.02081	-3.18666	0.034484	-1.70569	1.257464	-0.63974	-0.24698	1
50904	-0.27448	1.242853	2.246166	-1.32163	1.745691	0.445754	1.113257	0
51111	-2.51903	-1.82775	0.612049	-0.55001	-0.54029	0.273842	0.503184	1
51230	-1.06039	2.294035	-0.44217	-0.13719	0.844419	-0.70788	0.211432	3
51264	-1.93252	-1.78707	0.741528	-2.24237	0.771232	-1.81262	-0.81809	1

[군집 분석 결과]

PC1	구종 비율(CLR) + 구속/무브먼트 일부 기여도 높음
PC2	vx, vy, vz 등 속도 성분 기여가 큼
PC3	speed_std, az, z 위치 변수 영향 큼
PC4	x/z 위치 + 가속도 혼합
PC5~7	특정 구종 비율 또는 물리량이 부분적으로 작용

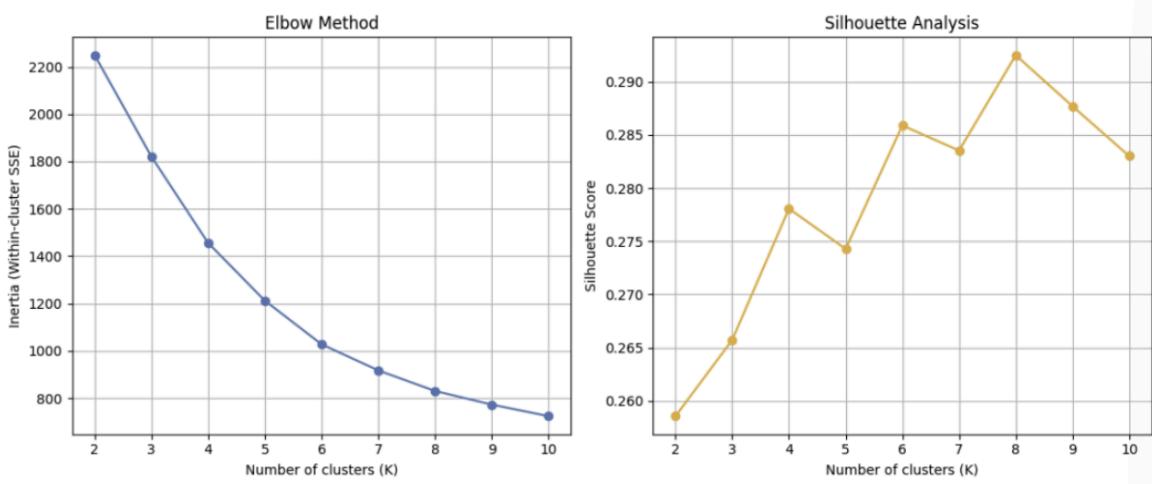
[주성분별 분석]

PCA4 clustering (PC1-PC2 projection)



[PC1-PC2 평면에서 PCA4 기반 군집 결과를 다시 시각화한 산점도]

추가적으로 구속/상하 무브먼트 중심, 좌우 움직임 중심, 변동성/결정구 성향, 공의 높낮이 조절과 제구 스타일 반영의 4가지 주성분을 바탕으로 군집분석을 진행하였다.



[군집 개수 결정]

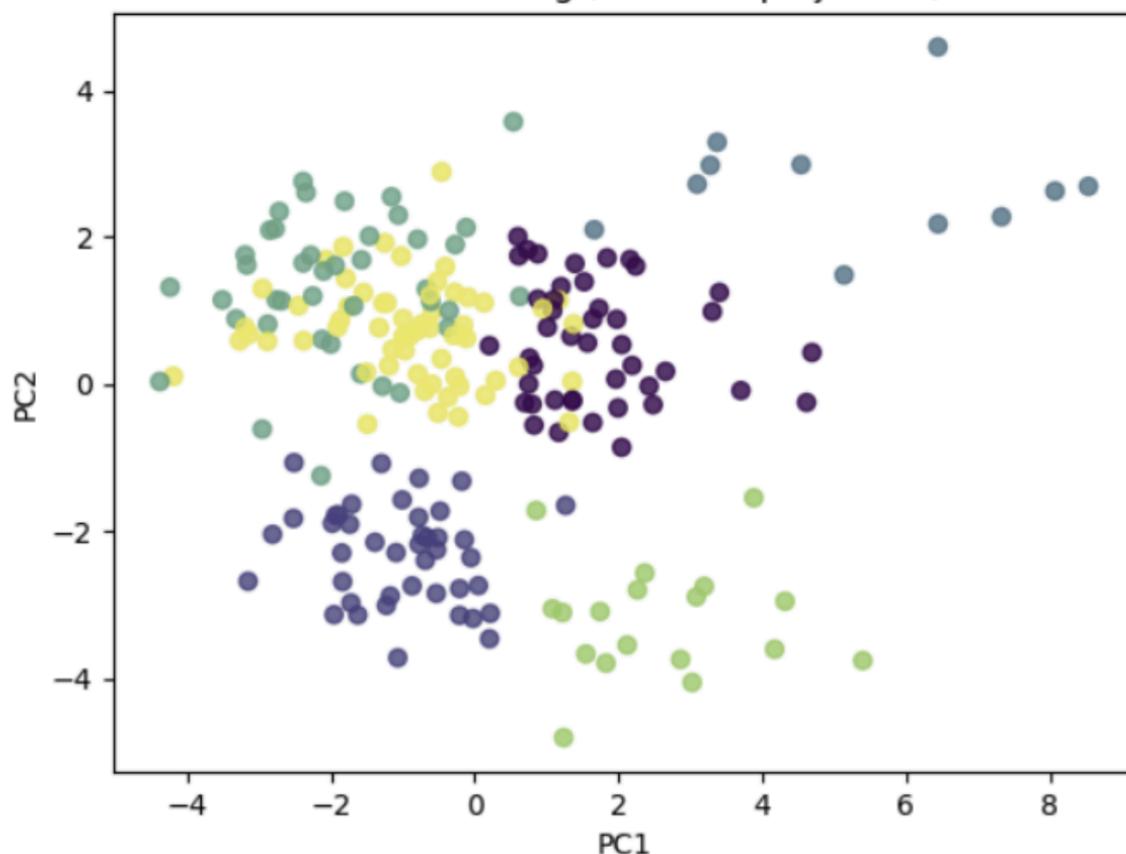
pitcherId	PC1	PC2	PC3	PC4	cluster
50030	3.704672	-0.08985	-0.83542	2.698763	0
50106	-2.71544	2.342735	-1.96556	0.990573	3
50109	0.05381	-2.74106	-1.0045	-2.31285	1
50354	-1.62459	-3.14009	-1.60456	-1.60243	1
50393	1.203678	1.322219	-0.06931	-0.37522	0
50441	0.878547	1.767839	-0.95061	0.604371	0
50464	-0.14173	-2.11825	-0.98652	-2.68994	1
50556	0.131538	1.104241	1.617899	0.729894	5
50662	-2.87657	0.811855	-0.98496	2.068991	3
50812	0.858416	-1.72024	-1.19866	0.575678	4
50859	-0.02081	-3.18666	0.034484	-1.70569	1
50904	-0.27448	1.242853	2.246166	-1.32163	5
51111	-2.51903	-1.82775	0.612049	-0.55001	1
51230	-1.06039	2.294035	-0.44217	-0.13719	3
51264	-1.93252	-1.78707	0.741528	-2.24237	1
51454	-0.20748	-3.14555	0.844605	0.302913	1
51516	-1.89631	-1.77864	0.398631	-1.67373	1
51594	-2.95286	-0.61461	-1.34581	-0.80486	3
51648	-1.95456	-3.13572	0.85443	-1.5419	1
51713	-0.67048	1.28798	-1.90037	1.448729	3
51715	-0.77127	-2.17923	0.804743	-2.08767	1

[군집 분석 결과]

PC1	Vz0_kmh_mean, z0_m_mean 등 구속/상하 무브먼트 중심
PC2	Vx0_kmh_mean, x0_m_mean 등 좌우 움직임 중심
PC3	speed_std, az_ms2_std, 커브 비율 등 변동성/결정구 성향
PC4	공의 높낮이 조절과 세부 제구 스타일 반영

[주성분별 분석]

PCA4 clustering (PC1-PC2 projection)



[PCA4 기반 군집 결과 시각화 산점도]

5.의의 및 한계점

본 연구는 기존의 단편적인 결과 데이터에 의존하던 전형적인 방법론에서 벗어나, 맥락 데이터와 물리 데이터를 적극적으로 반영함으로써 야구 분석의 새로운 방법론을 제시하였다라는 점에서 큰 의의를 지닌다. 특히 XGBoost 기반의 구종 예측 모델을 실질적으로 구현하여, 직전 투구와 타석 상황은 물론 이닝 및 경기 단위의 다각적인 맥락 정보를 투구 데이터 분석에 성공적으로 결합하였다. 이는 단순 통계 중심의 분석을 넘어 실질적인 경기 흐름을 예측 모델에 반영했다는 점에서 기술적 가치가 높다.

다만, 연구 과정에서 새로운 데이터 형식을 유연하게 처리하는 데 다소 미흡함이 있었으며, 활용된 변수(Feature)의 제약으로 인해 예측 모델의 성능을 극대화하는 데 일정 부분 한계가 존재했다. 이러한 한계를 극복하기 위해 향후에는 데이터 간의 복합적인 관계성을 파악할 수 있는 그래프 모델링(Graph Modeling)을 도입할 예정이다. 이를 통한 네트워크 분석 고도화는 기존 모델의 성능적 제약을 해소하고 더욱 정밀한 예측을 가능케 할 것으로 기대된다.