

paper proposal

# 통계모형 기반 투수 평가 지표 개선 방안



2025.5.23

# Table of Contents

01 개요

---

02 기존 투수 평가 지표

---

03 사용 모델

---

04 변수 선정 및 개선된 지표 제안

---

05 분석 결과 및 투수 평가

---

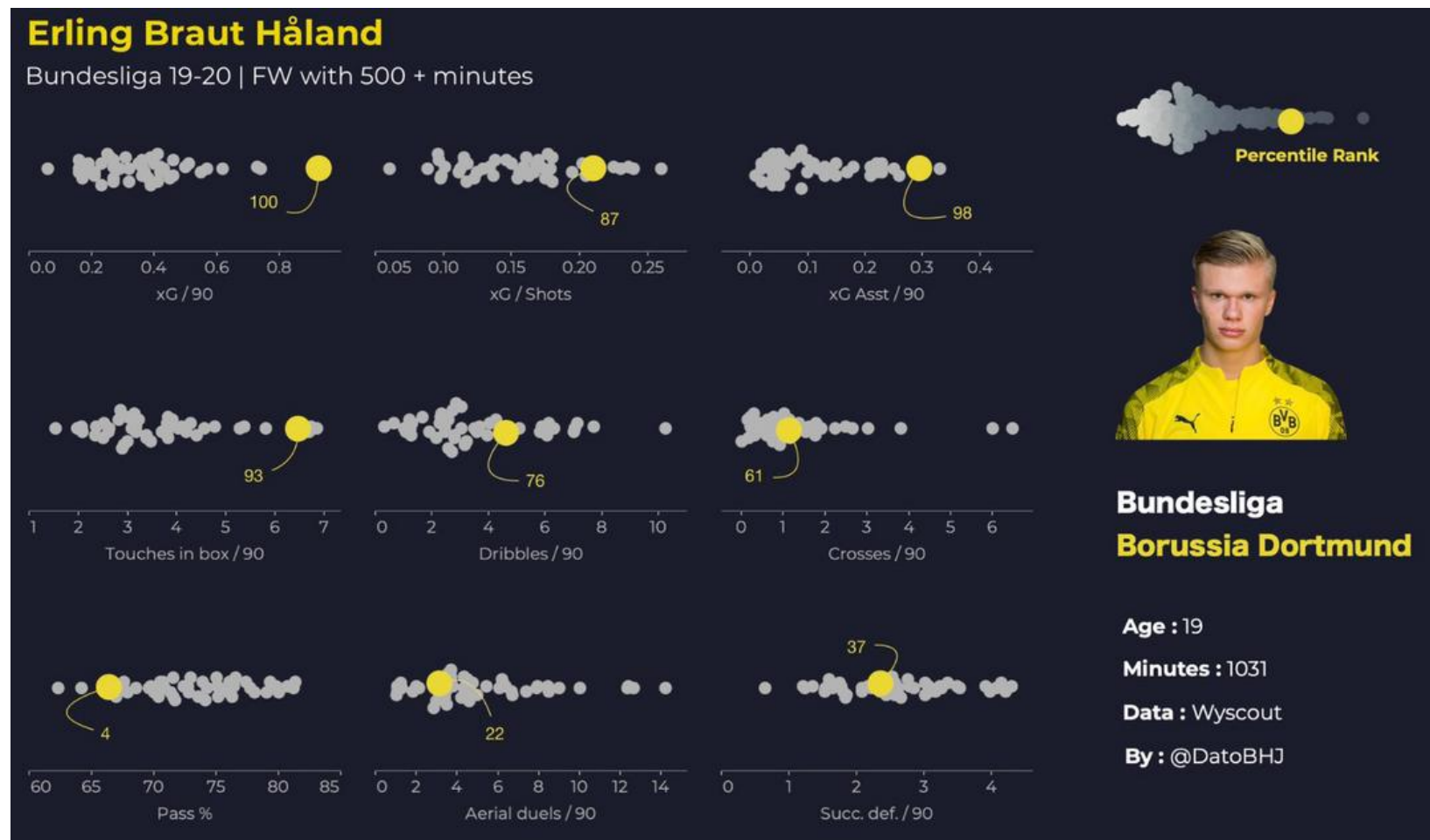
06 결론

---

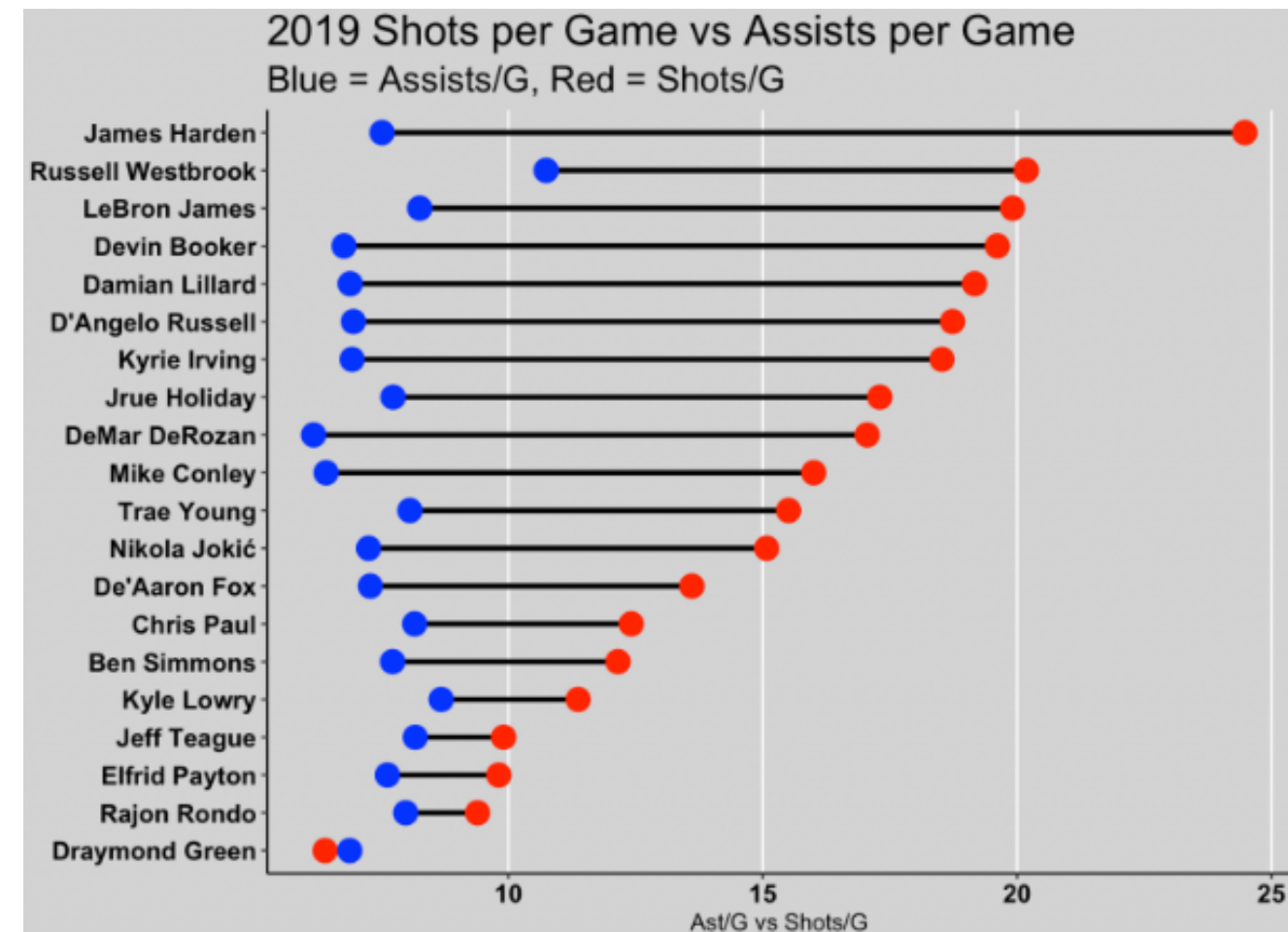
# 개요

최근 스포츠 분야에서 인공지능 모델에 기반해 선수들의 경기력을 향상시키는 방법과 도구는 다각도로 연구되고 있음  
→ 선수들의 경기력 측정을 위해 새로운 지표를 제안하는 선수들의 가치 평가 연구가 진행되고 있음

- 축구의 경우, 기대 득점(xG)과 기대 도움(xA) 등 선수 개인의 능력을 평가하는 수치화된 지표 개발
- 농구의 경우, 효율적인 득점 능력(EFG%)와 대체선수 평가기여도(PER) 등 수치화된 지표 개발



축구 지표 분석(출처 : 부산대학교)



농구 지표 분석(출처 : optimumsportsperformance)

# 개요

야구의 경우, 사람의 행동을 분석하는 영역 and 지표를 분석하는 연구가 이루어지고 있음

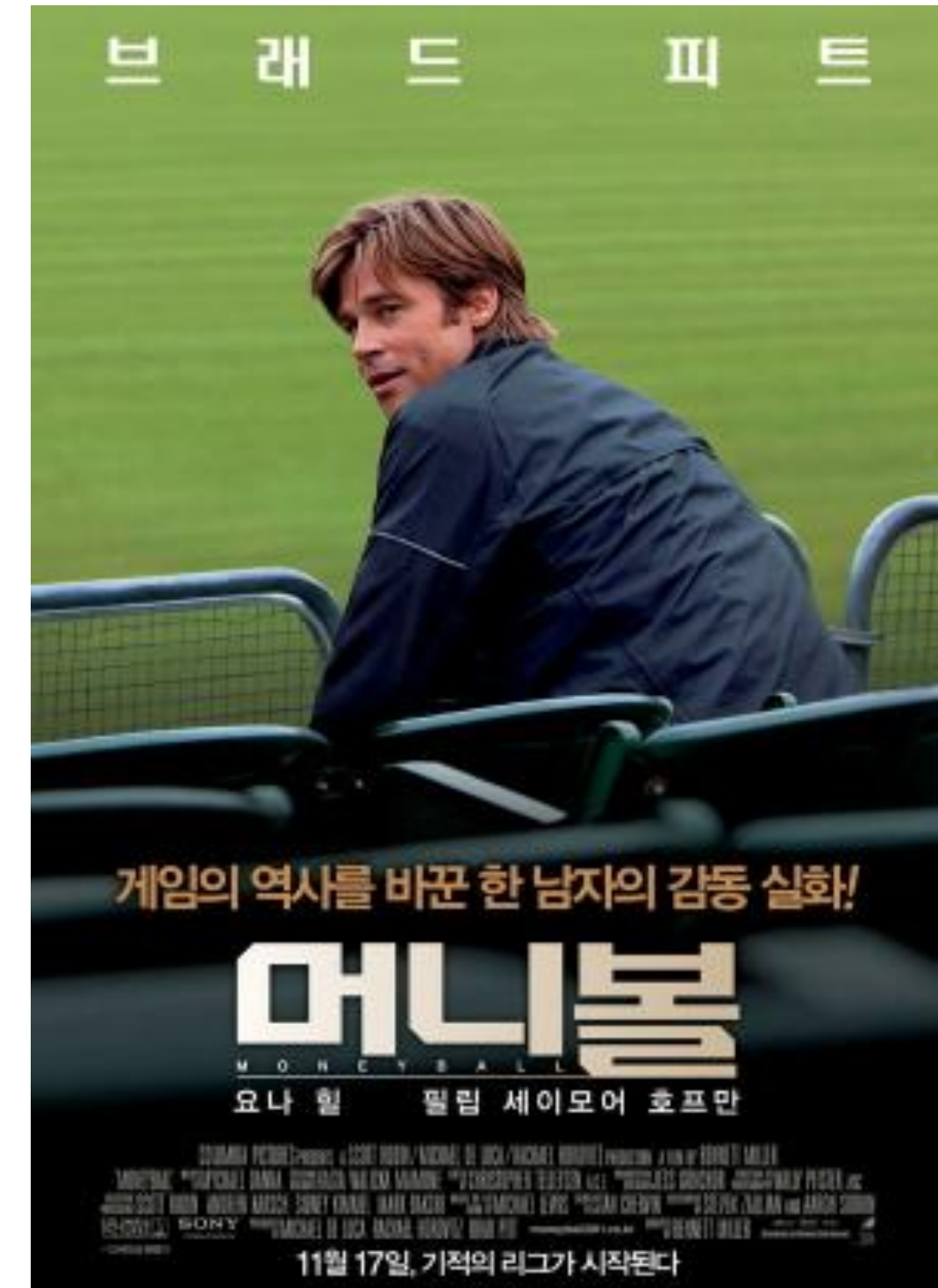
야구에서 기록된 경기 결과를 토대로 확률을 통해 통계적으로 분석하는 개발도구를 '세이버매트릭스'라고 함 (Baseball Abstract, 1977)[1]

1980년대부터 메이저리그(MLB)에서 도입, 2002년 오클랜드 어슬레틱스의 빌리 빈 단장이 팀 운영의 방침으로 세이버메트릭스를 도입

→ 성적향상으로 관심도가 커짐, 현재 구단운영의 핵심이 되었음

## 국내

- 타자능력지수와 투수능력지수를 제안(이제영, 2016) [2]
- 기계학습 모형을 적용하여 야구 기록을 분석 및 해석(주윤태, 2023) [3]
- 국내 구단은 팀마다 데이터 분석팀을 운영하여 시즌을 준비



영화 '머니볼' 포스터 (출처 : IDBM)

[1] The new Bill James historical baseball abstract

[2] 한국프로야구에서 타자능력지수 제안-대체선수대비승수 (WAR) 을 중심으로

[3] 기계학습을 이용한 한국 프로야구 신인 투수들의 유형화 및 잠재력 평가

# 개요

투수를 평가하기 위해 사용하는 대표적인 지표  
평균자책점(ERA)&수비무관 평균자책점(FIP)

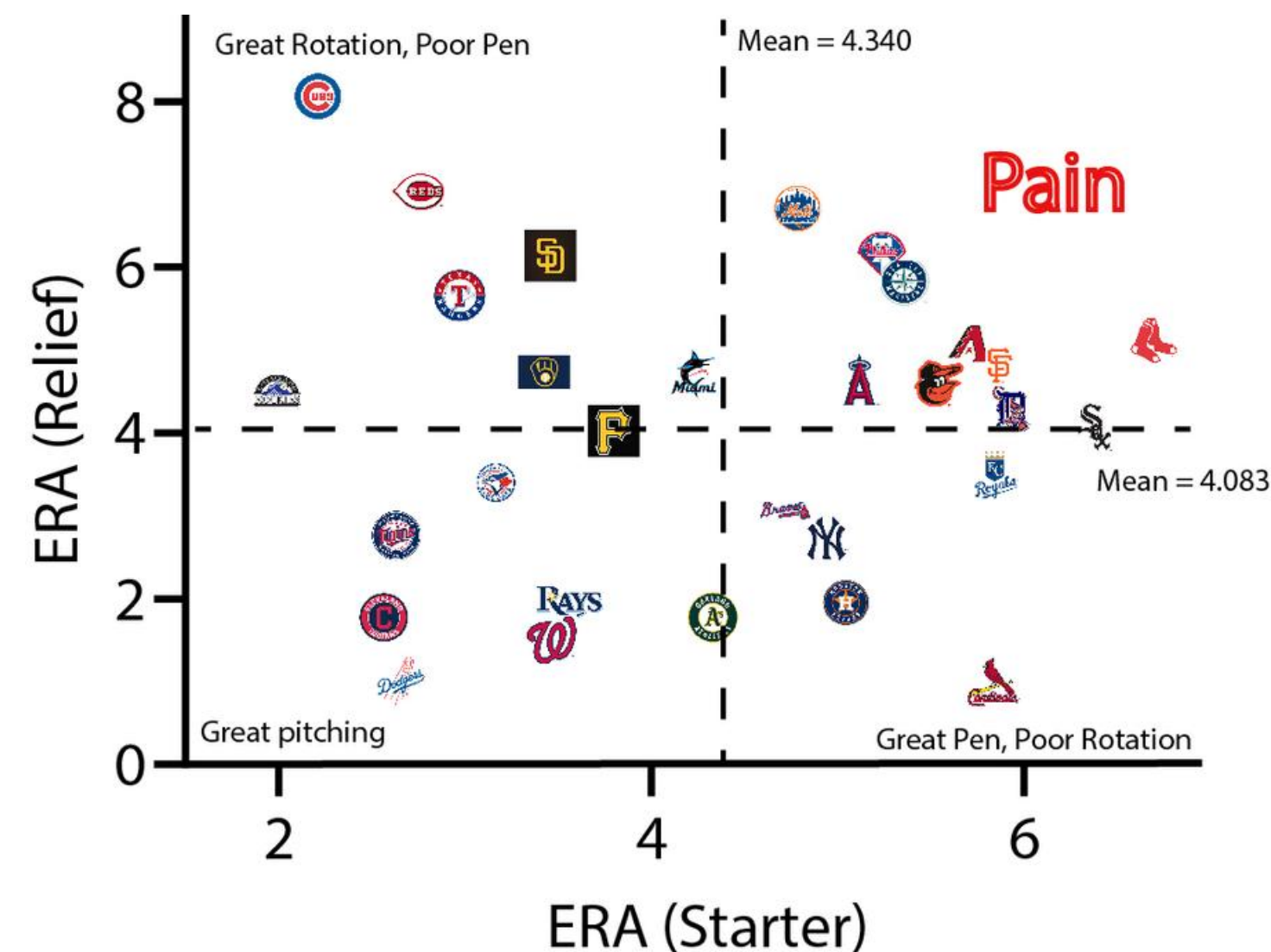
평균자책점 : 9이닝 당 투수가 허용한 실점을 고려하여 값을 산출

- 팀의 수비력에 영향을 많이 받는 지표  
(수비력이 약할수록, 타구를 처리하지 못하기 때문)

수비무관 평균자책점 : 투수의 삼진과 볼넷, 홈런을 고려하여 값을 산출

- 홈런 이외의 장타를 단타와 동일하게 평가  
(안타를 맞는 것은 투수의 책임이 많지 않음을 가정)

연구에서는 투수를 평가하는데 사용하는 지표의 단점을 개선  
→ 개선된 투수 지표를 제안



선발투수와 구원투수의 평균자책점 관계 (출처 : reddit)



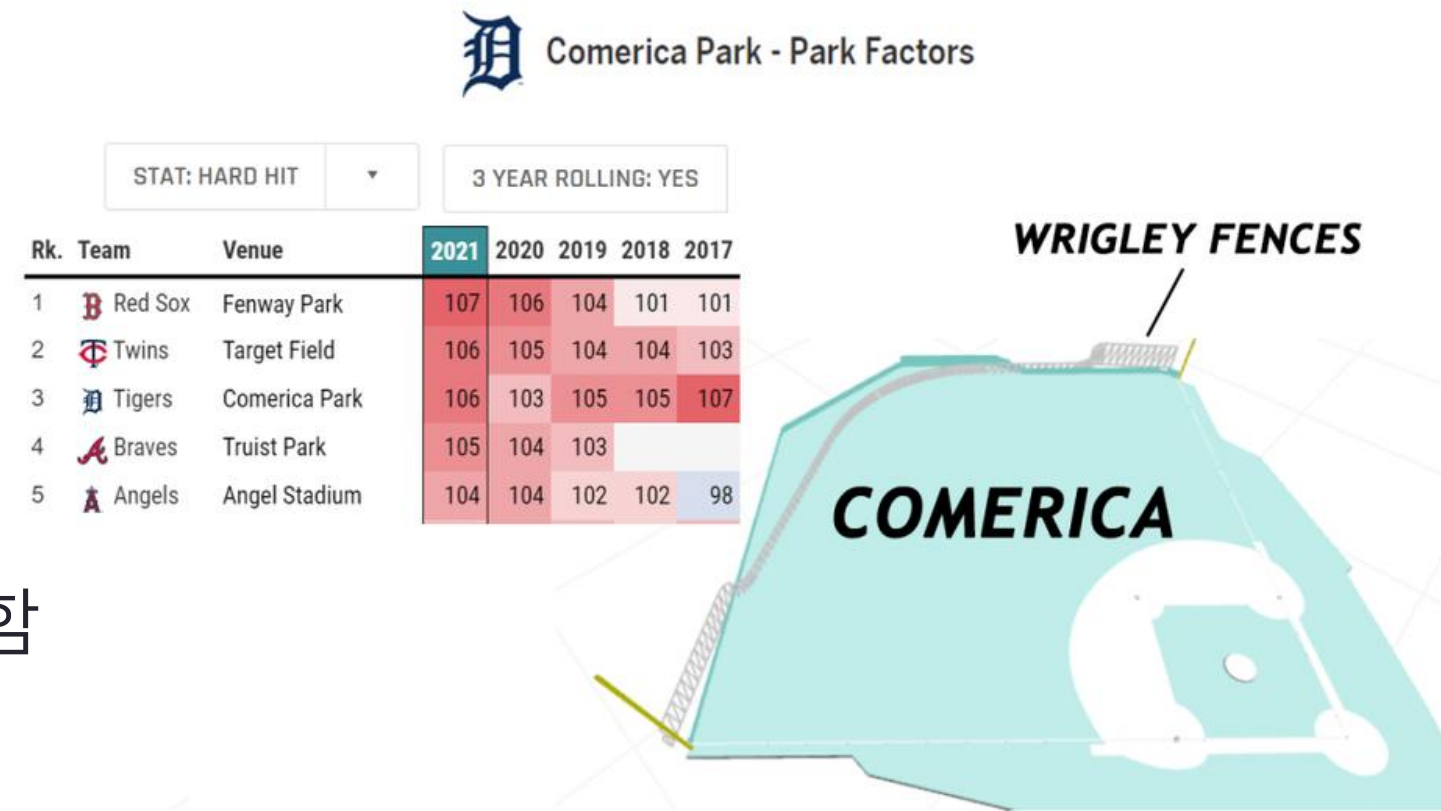
# 개요

개선된 투수 지표에 사용하는 적절한 구장 파크팩터를 선정하도록 제안

구장 파크팩터 : 구장에서 진행되는 경기 기록을 고려하여 타자&투수에게 어떤 영향을 미치는지 수치화하는 지표











연구에서는 투수를 평가하는데 있어 적절한 구장 파크팩터 산정식을 제안하도록 함

- 구장의 영향을 고려하여 투수의 실제 능력을 파악

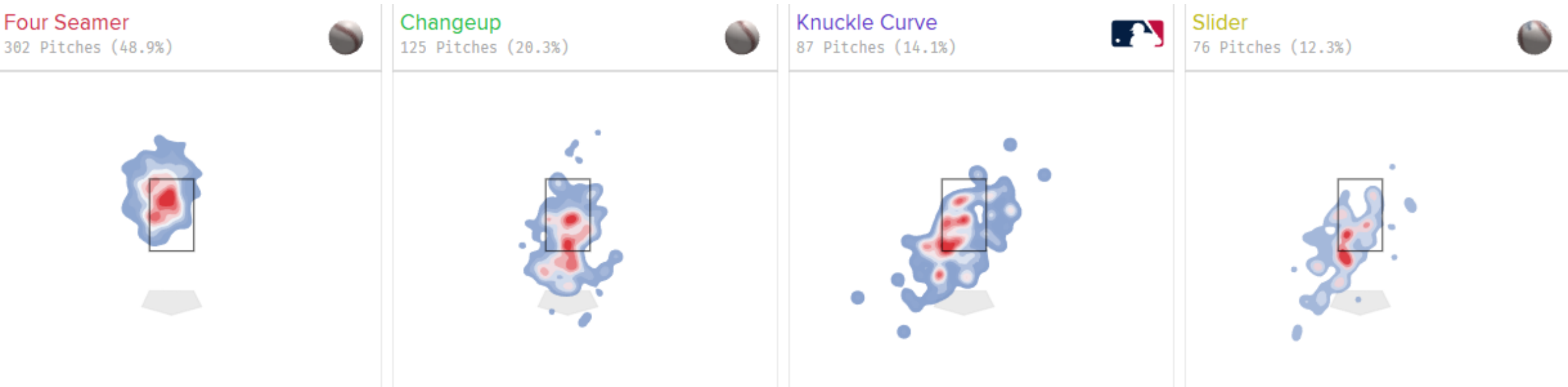


구장 파크팩터 설명 (출처 :MLB)

세이버메트릭스를 사용한 야구 연구 방향

Rk.	Player	Year	BF	K%	BB%	wOBA	xwOBA	LA Sweet-Spot %	Barrel%	Hard Hit %	EV50	Adjusted EV	Whiff %	Swing %
1	 Ragans, Cole	2025	148	38.5	7.4	.295	.249	30.4	8.9	38.0	76.3	94.2	34.5	49.3
2	 Gilbert, Logan	2025	117	37.6	5.1	.226	.250	28.8	4.5	40.9	79.1	94.2	38.3	51.3
3	 Wheeler, Zack	2025	198	33.3	4.5	.282	.253	37.2	8.3	35.5	77.3	93.6	31.7	50.5
4	 Greene, Hunter	2025	176	34.7	4.5	.250	.258	32.4	10.5	42.9	80.2	94.6	33.7	56.8
5	 Skenes, Paul	2025	190	24.7	6.3	.255	.260	28.5	4.6	38.5	77.6	93.6	27.0	50.1
6	 López, Pablo	2025	130	28.5	3.8	.239	.267	34.1	5.7	36.4	77.0	94.4	26.3	48.2
7	 Yamamoto, Yoshinobu	2025	175	30.3	8.0	.256	.268	28.0	5.6	40.2	78.0	94.6	31.7	45.8
8	 Mize, Casey	2025	167	21.0	5.4	.274	.269	29.3	8.9	40.7	79.8	94.3	28.3	49.5
9	 deGrom, Jacob	2025	149	24.2	6.0	.276	.269	26.2	6.8	33.0	75.3	93.3	30.8	50.8
10	 Liberatore, Matthew	2025	160	23.8	3.8	.245	.274	31.9	6.9	44.8	80.1	94.7	25.7	49.3

지표를 발전 및 창출하여  
투수의 가치를 평가하는 방법  
(출처 : statcast)



투수가 투구한 공을 분석하여  
투수의 구종 가치를 평가하는 방법  
(출처 : statcast)

# 기존의 투수 지표

평균자책점(ERA)은 투수가 9이닝을 기준으로 허용한 자책점을 의미

$$ERA = \frac{\text{자책점} \times 9}{\text{이닝}}$$

Ex. 한 투수가 9이닝 3실점을 기록, 평균자책점 3.00

## 장점

- 투수의 실력과 주자 억제력을 평가할 수 있는 좋은 지표
- 야구를 모르는 사람도 접하는 쉽고, 직관적인 지표

## 단점

- 야수진의 수비 능력에 따라 성적 편차가 크므로, 투수의 실력을 완벽히 평가하기 어려움



# 기존의 투수 지표

평균자책점(ERA)을 개선하기 위해 리그 평균의 평균자책점과 구장 파크팩터로 조정한 지표가 '조정평균자책점(ERA-)'

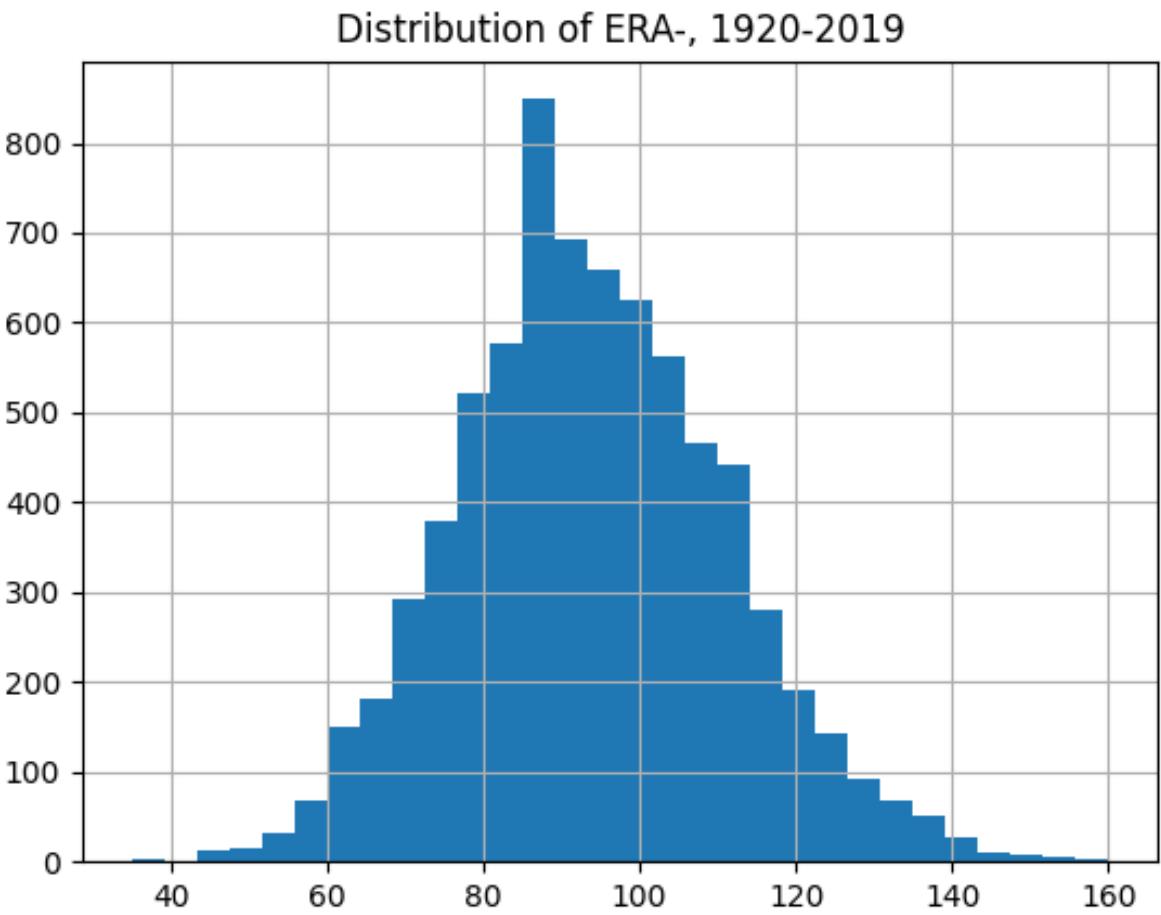
$$ERA- = 100 \times \frac{ERA + (ERA - (ERA \times PF))}{\text{리그 평균 ERA}}$$

### 장점

- 리그와 구장의 득점 수준에 관계없이 투수의 자책점이 얼마나 낮은지 확인 가능
- 상대 비교가 가능

### 단점

- 수비수의 책임을 무시함



역대 조정된 평균자책점 분포 (출처 :씩박꾸의 세이버메트릭스)



# 기존의 투수 지표

평균자책점의 단점인 야수진의 수비 능력과 팀 타선의 득점력이 요인으로 작용되는 문제를 해결할 필요가 있었음  
[Voros McCracken](#)은 홈런을 제외한 타구는 투수 평가에 전혀 영향을 미치지 않는다는 가정을 두고 식을 제안[4]

$$FIP = \frac{-2 \times \text{삼진} + 3 \times (\text{볼넷} + \text{몸에맞는볼}) + 13 \times \text{홈런}}{\text{이닝}} + C$$

Ex. B투수가 9이닝 동안 삼진 9개, 볼넷 3개를 내주고 홈런을 1개 허용할 경우 기록되는 FIP는  $0.44 + C$   
여기서 C는 시즌 상수로 보통 3.2~3.5 사이로 책정된다.

## 장점

- 투구의 퀄리티를 평균자책점보다 중점적으로 고려하여 계산

## 단점

- 구장을 고려하지 않음
- 타구 억제 능력을 고려하지 못함

# 기존의 투수 지표

수비무관 평균자책점의 문제를 해결하기 위해 '조정된 수비무관 평균자책점(FIP-)'이 제안됨

$$FIP- = 100 \times \frac{FIP}{\text{리그평균FIP}}$$

장점

- 투수 간 상대평가가 가능

단점

- 타구 억제 능력을 고려하지 못함

# 기존의 투수 지표

본 연구에서는 기존 투수 평가 지표의 단점을 개선한 지표를 제안

투수 실력을 보다 객관적으로 평가할 수 있는 지표를 생성

## 방법

- 조정된 평균자책점(ERA-)와 조정된 수비무관 평균자책점(FIP-) 각각에 영향을 주는 변수를 선정
- 영향을 주는 변수들의 중요도를 고려해 새로운 산정식을 제안

## 사용 모델

- 회귀분석 모델 : 릿지(Ridge) 회귀, 라쏘(Lasso) 회귀, 엘라스틱 넷(Elastic Net)
- 기계학습 모델 : 랜덤포레스트(Random Forest), XGBoost, LightGBM

## 평가 방법

- 기존의 평균자책점, 수비무관 평균자책점과 비교하여 성능이 좋은지 판단

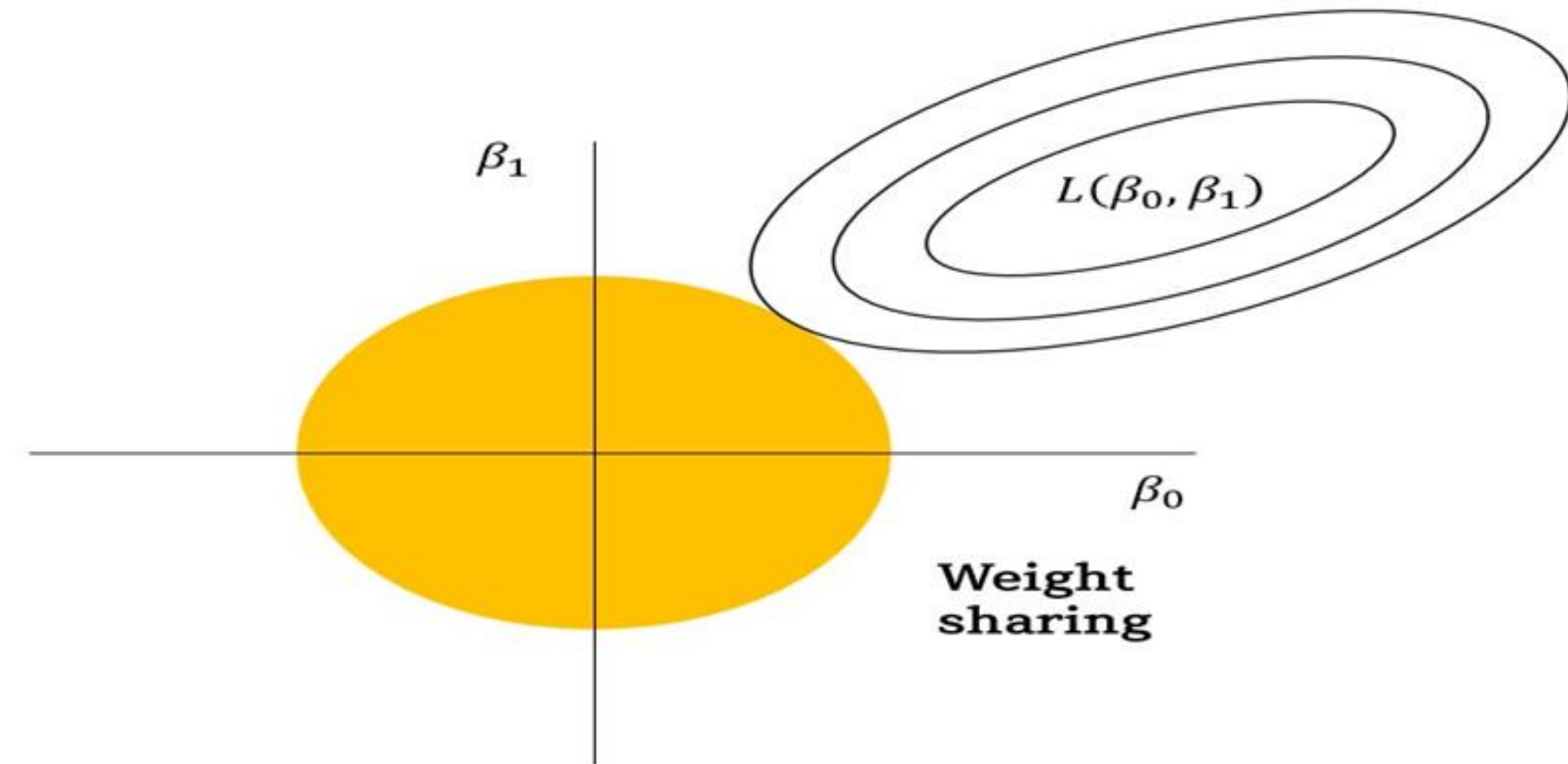
# 사용 모델

릿지 회귀(Ridge Regression)은 Hoerl, A.E와 Kennard, R.W가 제안한 모델[5]

기존 선형 회귀의 과적합을 해결하고 다중 선형 회귀에서 자주 발생하는 다중공선성 문제를 해결하기 위해 사용

변수 간 상관관계가 높을 때 예측 성능이 높아질 수 있음

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$



릿지 회귀의 제약 조건 영역과 손실 함수 등고선 시각화



# 사용 모델

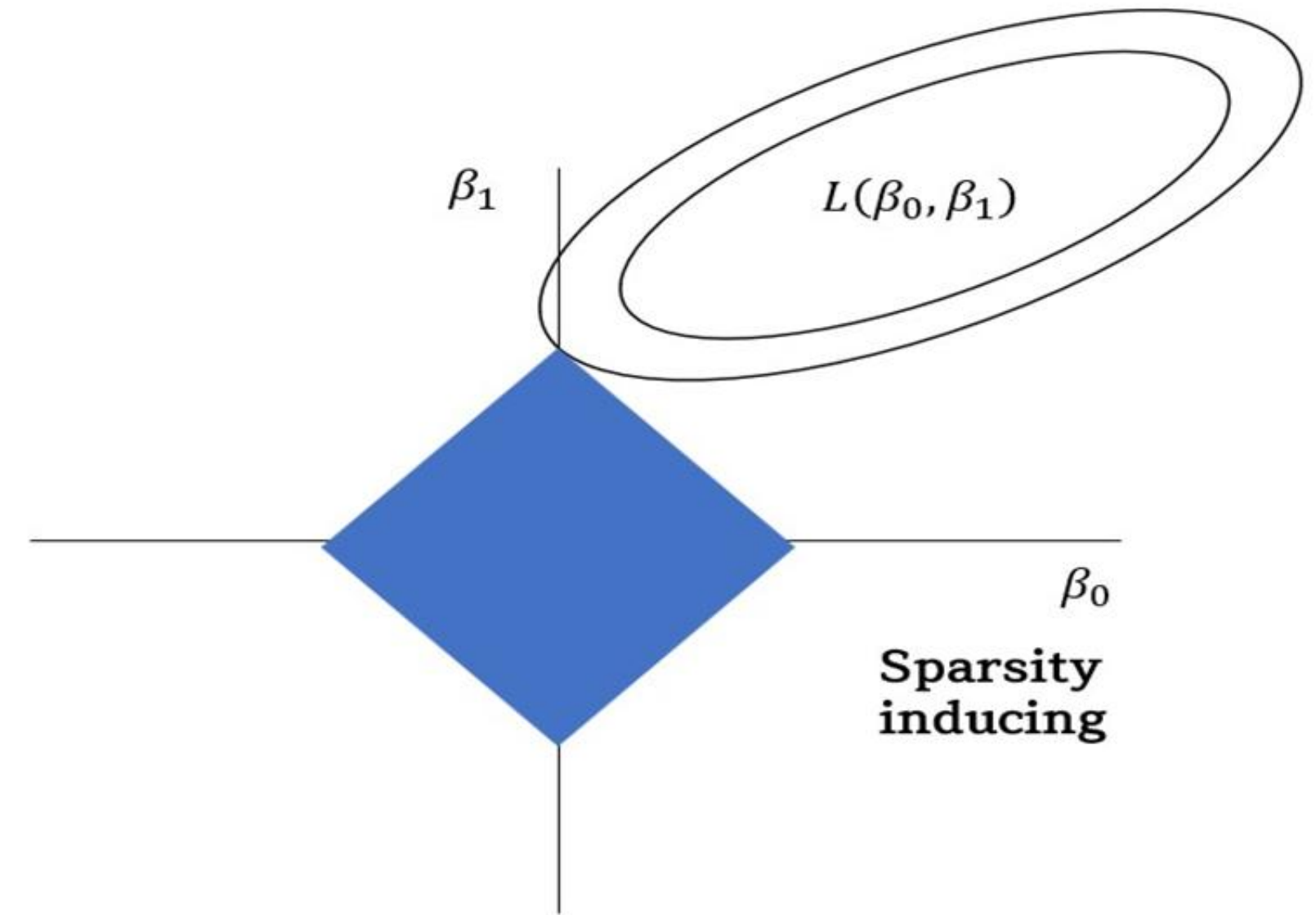
라쏘 회귀(Lasso Regression)은 Tibshirani, R이 제안한 모델[6]

제약조건을 만족하면서 오차가 최소화된 변수를 선택

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

$\lambda = 0$ , OLS 방식과 동일

$\lambda = \infty$ ,  $\beta$ 에 대해 제약이 큼



라쏘 회귀의 제약 조건 영역과 손실 함수 등고선 시각화

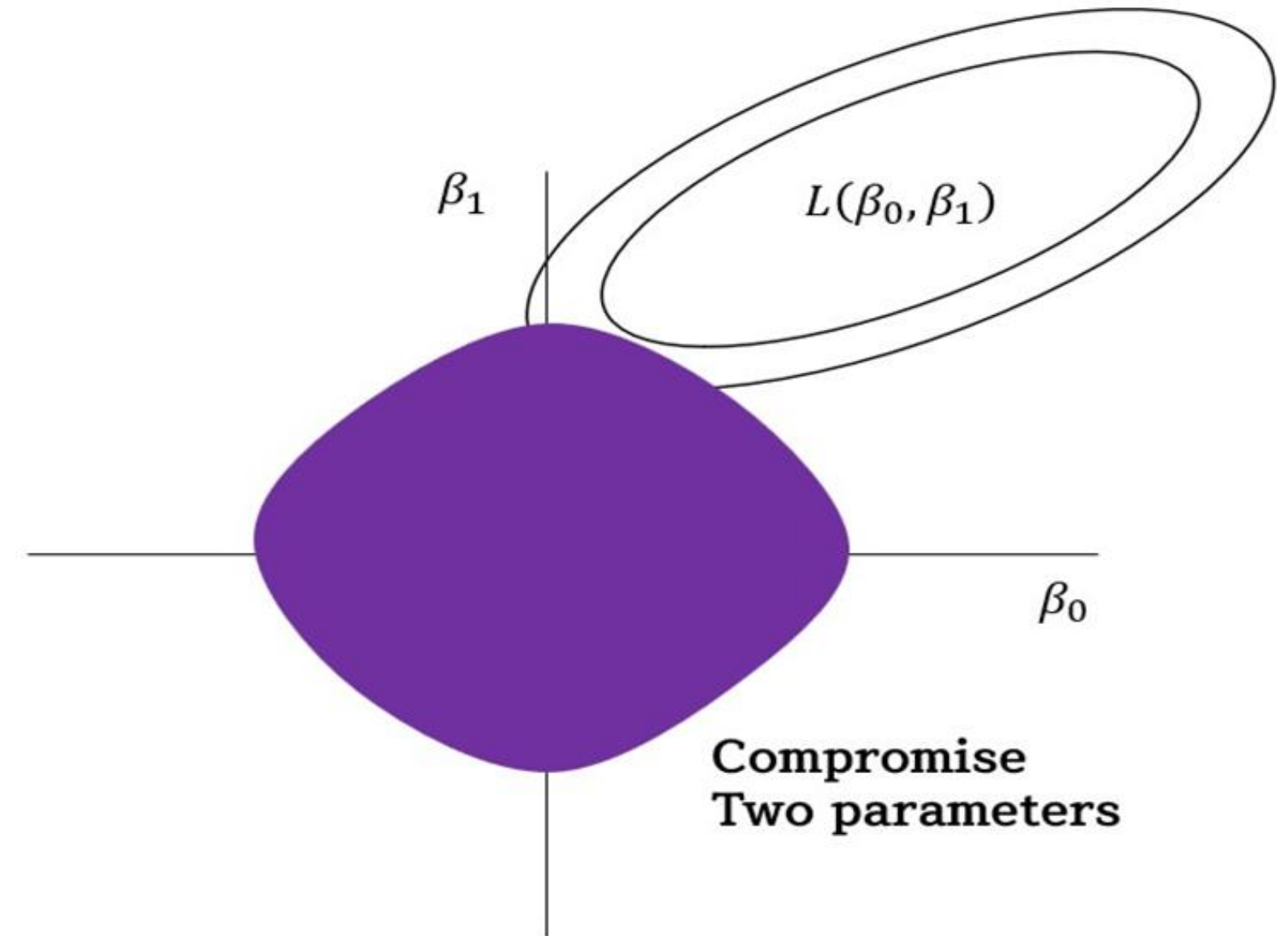
# 사용 모델

엘라스틱 넷(Elastic Net)은 Zou, H와 Hastie, T가 제안한 릿지 회귀와 라쏘 회귀를 복합적으로 활용하는 정규화 기법[7]

하이퍼파라미터  $\lambda$ 을 활용하여 각 기법의 영향도 조절 가능  
(한 가지 기법만 활용해도 가능)

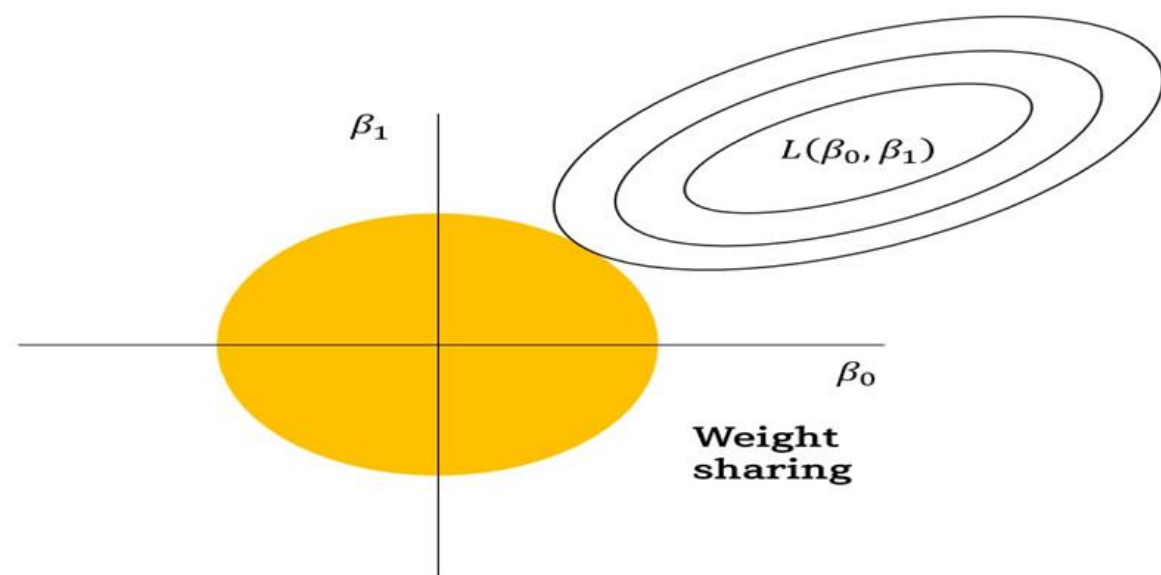
상관관계가 높은 변수들을 동시에 선택하는 특징

$$\hat{\beta}^{enet} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\}$$

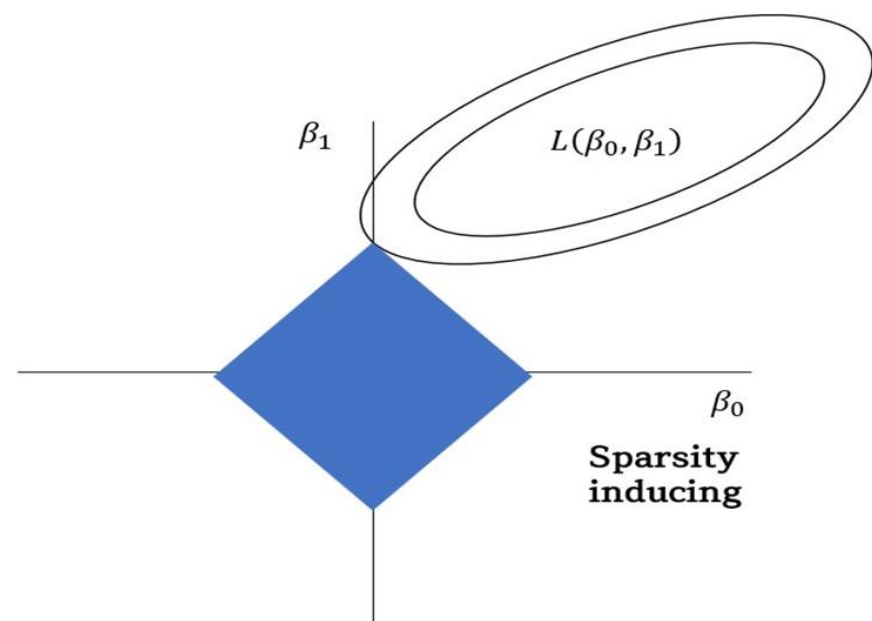


엘라스틱 넷의 제약 조건 영역과 손실 함수 등고선 시각화

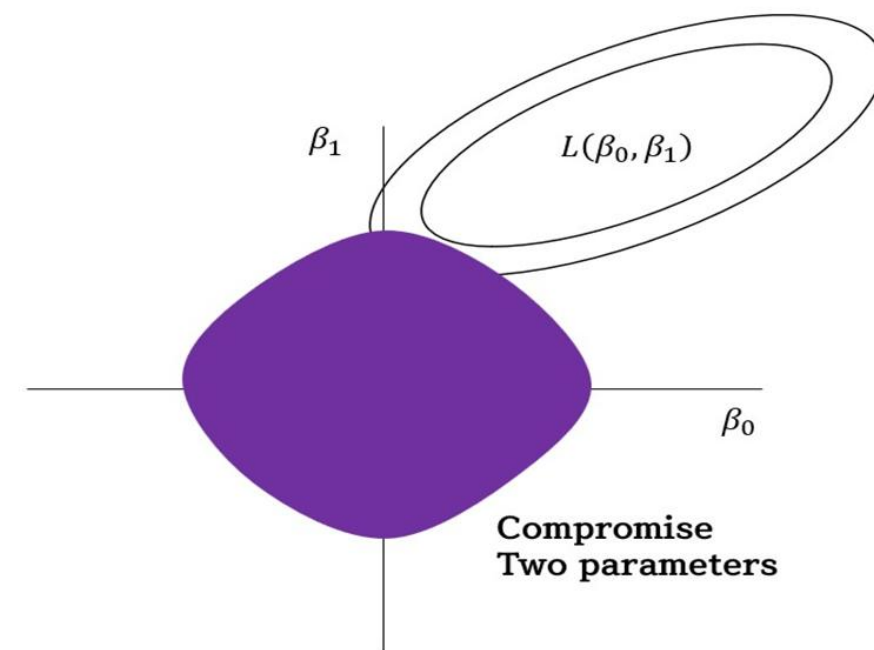
# 사용 모델



Ridge



Lasso



ElasticNet

분석을 위한 회귀 모델로 Ridge, Lasso, ElasticNet을 사용

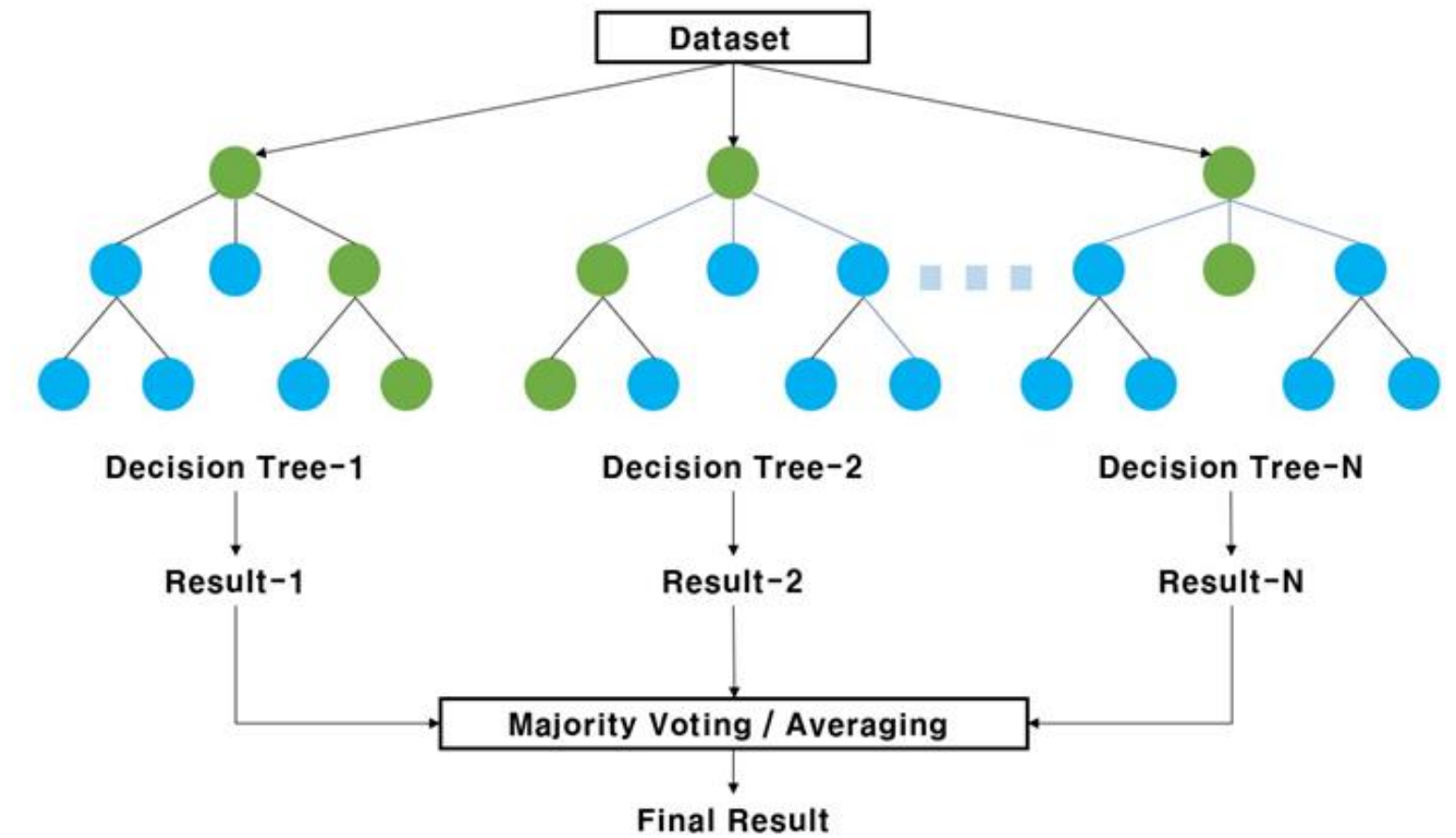
# 사용 모델

랜덤포레스트(Random Forest)는 Breiman, L이 제안한 모델[8]

의사 결정 나무의 일부 한계를 개선하도록  
앙상블(Ensemble method) 방법으로 결합한 모델

앙상블 방법을 사용하여 전반적으로 높은 정확도를 가짐

모델 학습 중 일어나는 과대적합을 방지하기 위해, 최적의 기준  
변수를 랜덤하게 선택하여 노이즈 데이터나 이상치가 모델에  
미치는 영향을 감소



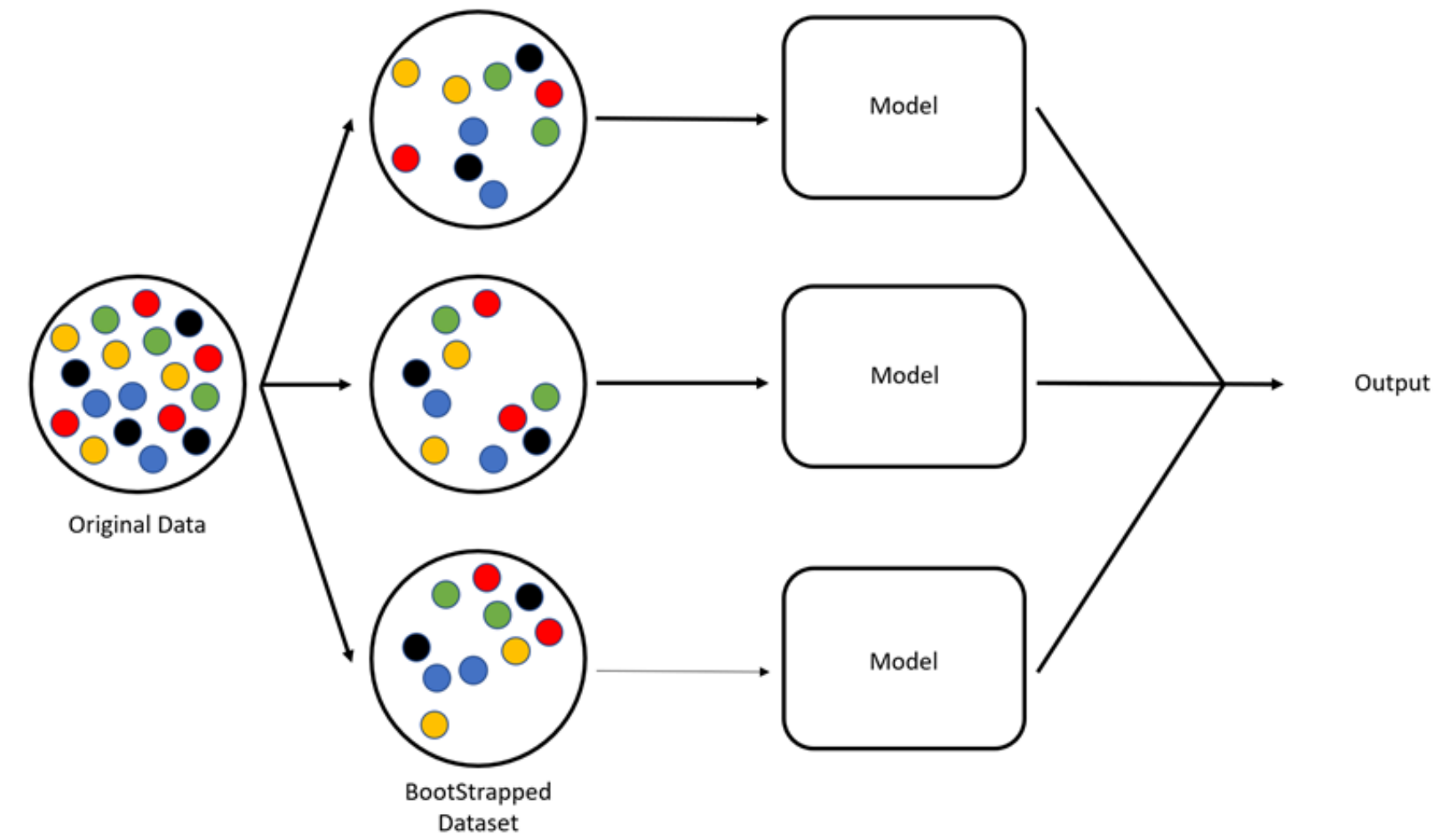
랜덤 포레스트의 모델 구조

# 사용 모델

랜덤포레스트(Random Forest)에서는 배깅(Bagging)을 사용

데이터셋을 독점으로 생산하는데 있어 복원추출을 사용, 붓스트랩에 복잡도가 높은 알고리즘을 개별적으로 학습해서 최종적으로 결합하므로 효과적

개인의 모델 성능은 떨어지더라도, 개별적으로 결합할 때 결과가 더 좋게 나오는 효과를 가짐  
→ 예측력 우수 및 변수의 중요도를 산출



배깅 방식 도식화



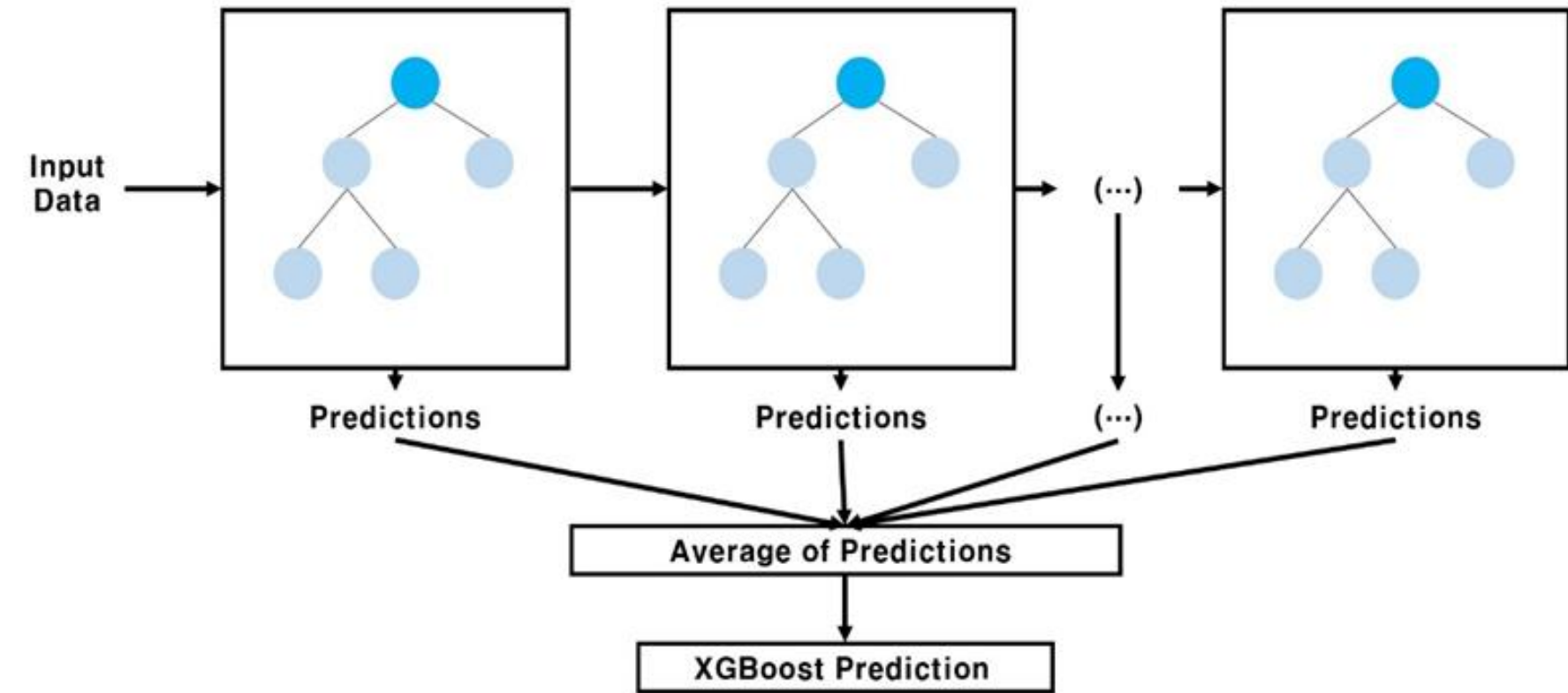
# 사용 모델

XGBoost는 Chen, T와 Guestrin, C가 제안한 모델[9]

그래디언트 부스팅(Gradient Boosting)을 사용하여 과적합  
방지를 고려한 기계학습 방법론

병렬처리를 통한 빠른 학습, 유연한 learning system,  
과대적합을 방지하기 위한 설계 등을 고려

트리 부스팅(Tree Boosting)을 진행하여 모델 적합 과정 중,  
편향과 분산의 반비례 관계(Trade off)를 고려해 예측 성능을  
높이는 과정



XGBoost 모델 구조

[9] Xgboost: A scalable tree boosting system

# 사용 모델

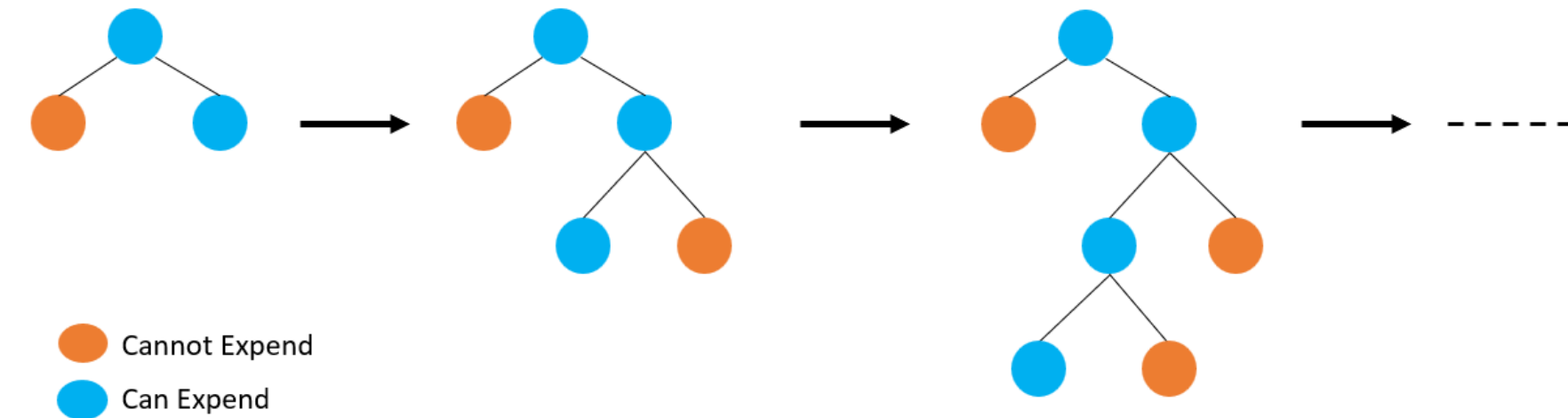
LightGBM은 KE, G가 제안한 모델[10]

모든 데이터를 스캔하지 않는 기계학습 방법론

큰 데이터세트에 적합하며, 다른 기계학습 방법론에 비해 적은 메모리를 차지한다는 장점을 가짐

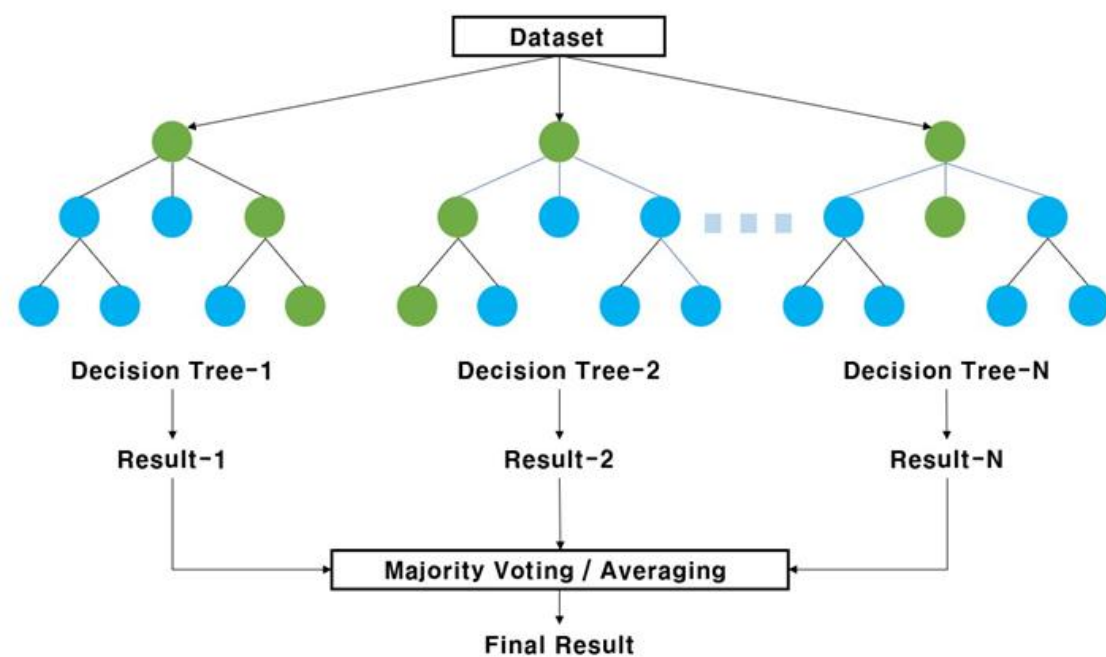
수직적으로 트리가 확장되는 구조

- 최대 손실 값을 가지는 리프 노트를 지속적으로 분할, 트리의 깊이가 깊어짐(비대칭적 트리 생성)
- 반복될수록 균형 트리의 분할 방식보다 예측 오류 손실을 최소화

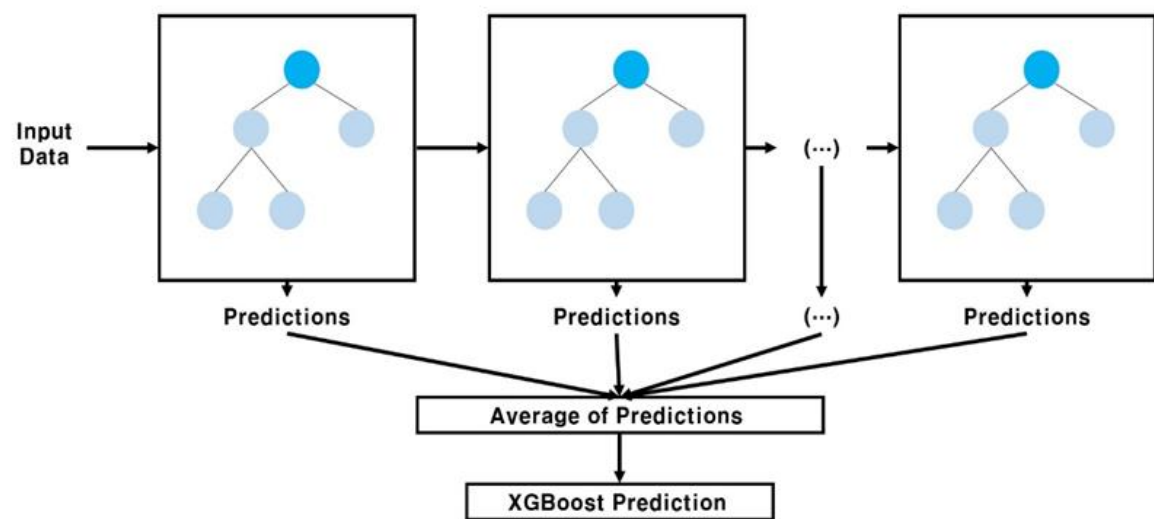


(Fig. LightGBM의 모델 구조)

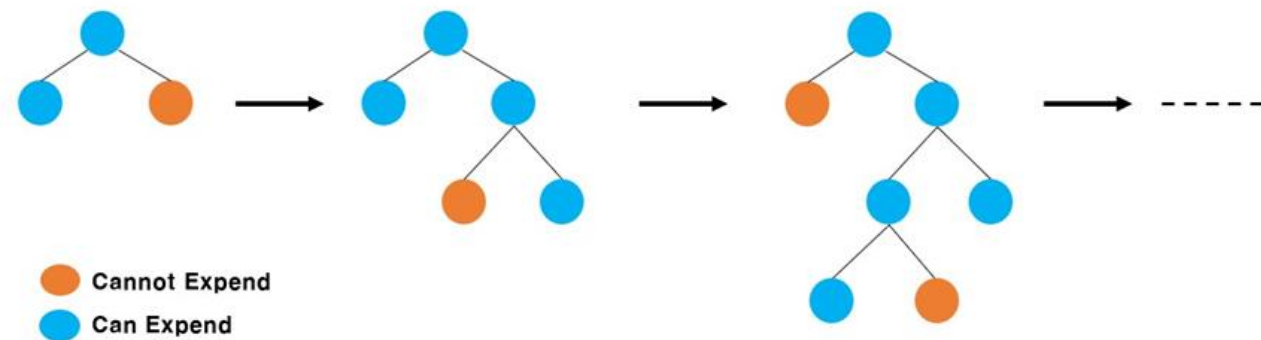
# 사용 모델



Random Forest



XGboost



LightGBM

분석을 위한 기계학습 모델로 Random Forest, XGboost, LightGBM을 사용

# 변수 선정 및 개선된 지표 제안

## 변수 설명

1차스탯  
볼, 아웃, 파울 등의 기록으로 수치적인 지표를 의미

지표	설명
IP	투수가 소화한 이닝
SO	투수의 삼진 개수
BB	투수의 볼넷 개수
HR, H	투수가 허용한 피홈런과 피안타

지표	설명
R	투수가 허용한 실점
ER	투수가 허용한 실점 중 책임져야 하는 실점
BB/9	9이닝 당 허용한 볼넷 개수
HR/9	9이닝 당 허용한 홈런 개수



# 변수 선정 및 개선된 지표 제안

## 변수 설명

2차스탯  
1차스탯을 바탕으로 한 통계치를 고안하여 만든 지표

지표	설명
ERA	9이닝 기준 허용한 평균자책점
FIP	삼진, 볼넷, 홈런만으로 평가하는 평균자책점
ERA-	ERA에 시대적 보정과 구장 보정을 반영한 지표
FIP-	FIP에 시대적 보정을 반영한 지표

지표	설명
OBP, SLG	피출루율, 피장타율
BABIP	타석대비 인플레이 타구의 피안타율
RA9	9이닝 당 투수가 허용한 모든 실점
WAR	대체 선수 대비 팀 승리에 공헌한 승리수를 종합하여 표현한 지표





# 변수 선정 및 개선된 지표 제안

## 변수 설명

새로운 지표를 생성하는 과정 중 고려하는 변수 : '구장 파크팩터'  
'구장 파크팩터'란 각 구장의 성향을 나타내어 구장마다 타자와 투수 별 유리불리를 알 수 있는 지표

- 1을 기준으로 1보다 크면 타자에게 유리
- 1보다 작으면 투수에게 유리

예시로 경상국립대구장 파크팩터가 1.1인 경우, 중립구장에 비해 득점이 10% 더 나올 것으로 해석

$$PF(Parkfactor) = 100 \times \left( \frac{\frac{homeRS + homeRA}{homeG}}{\frac{roadRS + roadRA}{roadG}} \right)$$

지표	설명
homeRS	홈 득점
homeRA	홈 실점
homeG	홈 경기 수
roadRS	원정 득점
roadRA	원정 실점
roadG	원정 경기 수



# 변수 선정 및 개선된 지표 제안

## 변수 선정

### 분석 집단 기준 선정

분석 데이터 : 2020~2024시즌 (5년 동안 KBO리그 경기 기준)

분석 대상 : 상대한 타자 수가 100타석 이상인 투수들의 개인 데이터만 추출하여 분석[11]

→ 상대한 타자의 수 표본이 적은 경우는 노이즈를 줄 것으로 예상

### 분석 집단 구분

선발투수 : 0:0의 상황에서 경기를 시작하는 투구를 진행

구원투수 : 선발투수가 강판된 후, 경기 중간 or 마무리로 투구를 진행

→ 투수의 보직에 따라 투구 환경은 달라지며, 분석의 방향을 달리해야 할 것으로 고려

대표적으로 많이 사용하고 있는 'WAR(대체선수대비 승리기여도)' 지표도 선발투수와 구원투수의 환산식이 다름

### 분석

1. 선발투수의 평균자책점과 수비무관 평균자책점

2. 구원투수의 평균자책점과 수비무관 평균자책점

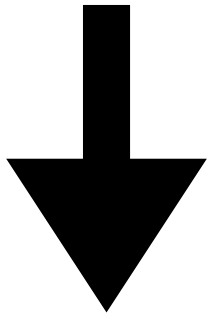
- 선발투수 308명, 구원투수 548명의 데이터를 가지고 분석

[11] 스카우트 리포트 2018, 레이더는 타구의 질을 어떻게 평가할까?

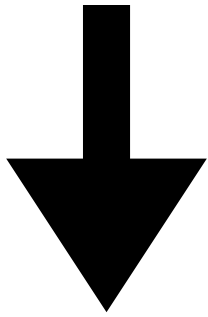
# 변수 선정 및 개선된 지표 제안

## 변수 선정

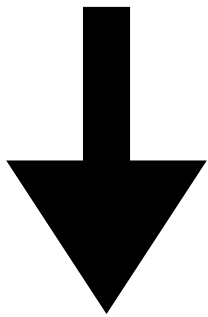
1. 종속변수 간 양의 상관관계가 0.5 이상인 변수를 추출



2. 단계적 선택법을 사용하여 최적의 변수를 출력할 수 있도록 모델 학습

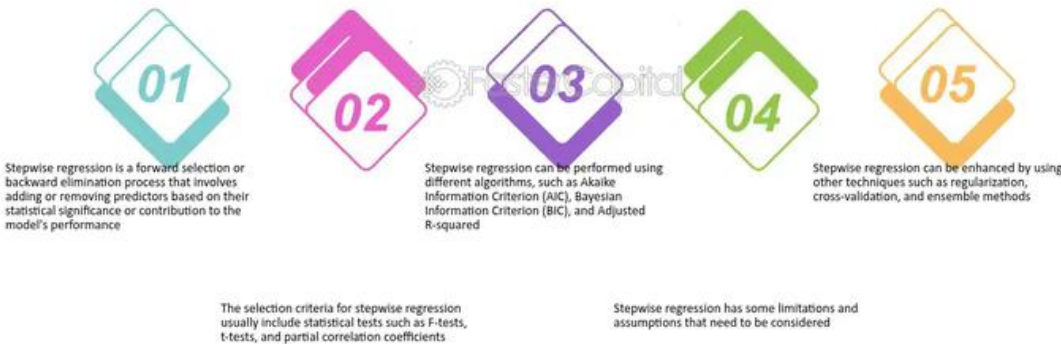


3. 모델을 고려할 때, AIC가 가장 낮은 변수 모델을 채택하여 이후의 분석을 진행[12]

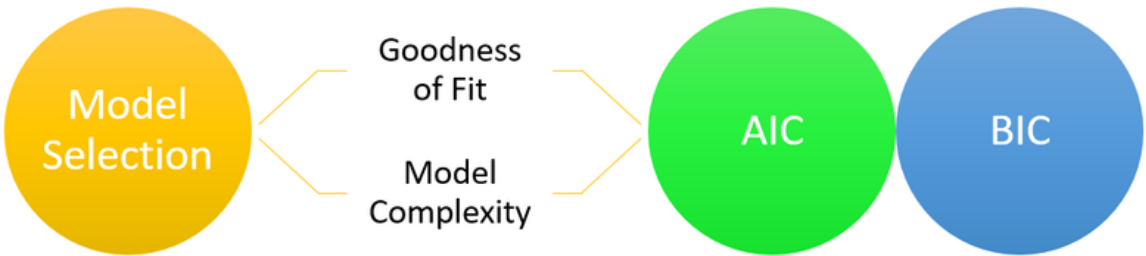


4. AIC가 가장 낮은 모델의 변수들이 다중공선성을 위배하지는 않는지 확인

## Introduction to Stepwise Regression



단계적 선택법 방식 (출처 : FasterCapital)



AIC 고려 변수 모델 선정 기준 (출처 : Allen's 데이터 맛집)

## How to Interpret VIF Values



다중공선성 설명 그림 (출처 : FasterCapital)

[12] Information theory and an extension of the maximum likelihood principle

# 변수 선정 및 개선된 지표 제안

## 변수 선정

분석대상	구분	변수	변수설명	양의 상관계수	AIC
선발투수의 조정된 평균자책점	종속변수	ERA-	조정된 평균자책점		2056.77
	독립변수	RA9	9이닝 당 실점	0.94	
		SLG	피장타율	0.78	
		K%	삼진율	0.56	
		BB/9	9이닝 당 볼넷 수	0.53	

AIC가 가장 낮은 최적의 모형으로 선정된 선발투수의 조정된 평균자책점에서 유의한 변수는 RA9, SLG, K%, BB/9가 선정되었음



# 변수 선정 및 개선된 지표 제안

## 변수 선정

분석대상	구분	변수	변수설명	양의 상관계수	AIC
구원투수의 조정된 평균자책점	종속변수	ERA-	조정된 평균자책점		4071.05
	독립변수	RA9	9이닝 당 실점	0.94	
		LOB	잔루율	0.75	
		SLG	피장타율	0.73	

AIC가 가장 낮은 최적의 모형으로 선정된 구원투수의 조정된 평균자책점에서 유의한 변수는 RA9, LOB, SLG가 선정되었음





# 변수 선정 및 개선된 지표 제안

## 변수 선정

분석대상	구분	변수	변수설명	양의 상관계수	AIC
선발투수의 조정된 수비무관 평균자책점	종속변수	FIP-	조정된 수비무관 평균자책점		1587.41
	독립변수	OBP	피출루율	0.76	
		HR/9	9이닝 당 홈런 수	0.7	
		BB/9	9이닝 당 볼넷 수	0.63	
		K/9	9이닝 당 삼진 수	0.62	

AIC가 가장 낮은 최적의 모형으로 선정된 선발투수의 조정된 수비무관 평균자책점에서 유의한 변수는 OBP, HR/9, BB/9, K/9이 선정되었음



# 변수 선정 및 개선된 지표 제안

## 변수 선정

분석대상	구분	변수	변수설명	양의 상관계수	AIC
구원투수의 조정된 수비무관 평균자책점	종속변수	FIP-	조정된 수비무관 평균자책점		3457.83
	독립변수	HR/9	9이닝 당 홈런 수	0.64	
		K%	삼진율	0.62	
		OBP	피출루율	0.61	
		SLG	피장타율	0.58	
		K/BB	삼진/볼넷 비율	0.53	

AIC가 가장 낮은 최적의 모형으로 선정된 구원투수의 조정된 수비무관 평균자책점에서 유의한 변수는 HR/9, K%, OBP, SLG, K/BB가 선정되었음



# 변수 선정 및 개선된 지표 제안

## 구장 파크팩터 선정

### 구장 파크팩터를 계산하는 개선 방식

- Acharya는 ANOVA 기반의 고정 효과 모델을 제안[13]
- Eiji Konaka는 로지스틱 회귀를 통해 파크팩터를 추정하는 방법을 제안[14]

### 본 연구에서 사용할 구장 파크팩터 식

본 연구에서는 MLB리그에 사용되고 있는 Baseball reference의 투수 구장 파크팩터 방식을 사용[15]

→ 현재 많은 분석 사이트의 파크팩터 산정식의 기틀이 된 방법

- 비교하는 방식 : 1년치 ~ 3년치 팀 데이터 중 무엇으로 구장 파크팩터를 산출하는 것이 적합한지

1년치 : 가장 최신 시즌의 환경(예: 구장 변경, 공인구 변화, 리그 환경 변화 등)을 반영

2년치 : 한 해의 특이치가 전체 PF를 뒤흔들 가능성을 완화

3년치 : 샘플 사이즈가 충분해 평균적인 구장 특성을 잘 반영

---

[13] Improving major league baseball park factor estimates

[14] Park factor estimation improvement using pairwise comparison method

[15] [baseball-reference.com/about/parkadjust.shtml](http://baseball-reference.com/about/parkadjust.shtml)

# 변수 선정 및 개선된 지표 제안

## 구장 파크팩터 선정

지표	설명	지표	설명	지표	설명	지표	설명
Home Win	홈경기 승리 수	Visit Win	원정경기 승리 수	Home get score	홈팀 홈경기 득점 수	Visit get score	홈팀 원정경기 득점 수
Home Loss	홈경기 패배 수	Visit Loss	원정경기 패배 수	Home give score	원정팀 홈경기 득점 수	Visit give score	원정팀 원정경기 득점 수
Home Draw	홈경기 무승부 수	Visit Draw	원정경기 무승부 수	Home Game	홈경기 수	Visit game	원정경기 수
Home win% with draw	무승부 포함 홈경기 승률	Visit win% with draw	무승부 포함 원정경기 승률	League score	리그 전체 득점 수	League game	리그 전체 경기 수



# 변수 선정 및 개선된 지표 제안

## 구장 파크팩터 선정

비교결과, 3년치 데이터를 사용하였을 때 결정계수가 제일 높고, RMSE가 제일 낮음을 확인

- 본 연구에서는 3년치 데이터를 사용해 앞으로 개선된 투수 지표를 생성하기 위한 분석에 활용

식	결정계수	RMSE
1년치 데이터 사용	0.9834	0.008
2년치 데이터 사용	0.8516	0.016
3년치 데이터 사용	0.9956	0.002



# 변수 선정 및 개선된 지표 제안

## 개선된 투수 지표 제안식

산정된 식에 구장 파크팩터를 고려하여 개선된 투수 지표를 계산

- 구장 파크팩터는 수치가 클수록, 투수에게 불리하기 때문에 역수를 취하여 계산하도록 함

## 회귀분석 모델에서 제안하는 산정식

$x$ 는 선정된 변수들을 의미하며,  $\beta$ 는 각 변수의 표준화 회귀계수

표준화 회귀계수로 개선된 지표 :  $(B_0^* + B_1^*x_1 + B_2^*x_2 + \dots B_n^*x_n) \times \frac{1}{\text{구장 파크팩터(PF)}}$

## 기계학습 모델에서 제안하는 방향

기계학습 모델은 변수 중요도에 따른 산정식을 제안할 수 없음

각 모델의 변수 중요도에 따라 어떤 변수를 더 많이 사용하였는지를 확인, 구장 파크팩터를 함께 고려

# 분석 결과 및 투수 평가

## 변수 영향력 측정

평균자책점과 수비무관 평균자책점을 개선하기 위해 page 28~31에서 선택된 변수들을 고려

### 사용 모델

회귀분석 모델 : 릿지 회귀, 라쏘 회귀, 엘리스틱 넷

기계학습 모델 : 랜덤포레스트, XGBoost, LightGBM

### 평가 방법

양의 상관계수, 결정계수, RMSE를 고려하여 가장 높은 설명력으로 투수 가치를 설명할 수 있는 모델 선택

회귀분석 모델에서는 '표준화 회귀계수'를 고려하여 상대적 영향도를 분석

기계학습 모델에서는 '변수 중요도'를 고려하여 예측에 기여한 정도를 측정



# 분석 결과 및 투수 평가

## 변수 영향력 측정(회귀분석 모델)

분석대상	변수	Ridge	Lasso	ElasticNet
선발투수의 조정된 평균자책점	9이닝 당 실점	27.95	28.52	28.52
	9이닝 당 볼넷 수	1.96	1.77	1.77
	피장타율	1.04	0.64	0.64
	삼진율	-2.62	-2.59	-2.59

분석대상	변수	Ridge	Lasso	ElasticNet
선발투수의 조정된 수비무관 평균자책점	9이닝 당 볼넷	10.47	10.57	10.57
	9이닝 당 홈런 수	9.5	9.55	9.55
	피출루율	1.37	1.25	1.25
	9이닝 당 삼진 수	-10.53	-10.6	-10.6

# 분석 결과 및 투수 평가

## 변수 영향력 측정(회귀분석 모델)

분석대상	변수	Ridge	Lasso	ElasticNet
구원투수의 조정된 평균자책점	9이닝 당 실점	34.19	34.16	34.16
	피장타율	2.34	2.06	2.06
	잔루율	-0.12	0	0

분석대상	변수	Ridge	Lasso	ElasticNet
구원투수의 조정된 수비무관 평균자책점	9이닝 당 홈런 수	23.9	23.9	23.9
	피출루율	14.22	14.22	14.22
	삼진/볼넷 비율	-1.74	-1.74	-1.74
	삼진율	-12.19	-12.19	-12.19
	피장타율	-18.51	-18.51	-18.51

# 분석 결과 및 투수 평가

## 변수 영향력 측정(기계학습 모델)

분석대상	변수	XGBoost	Random Forest	LightGBM
선발투수의 조정된 평균자책점	9이닝 당 실점	76.97%	97.64%	37.66%
	9이닝 당 볼넷 수	6.57%	1%	22.63%
	피장타율	9.09%	0.78%	21.64%
	삼진율	7.36%	0.59%	18.07%

분석대상	변수	XGBoost	Random Forest	LightGBM
선발투수의 조정된 수비무관 평균자책점	9이닝 당 볼넷	14.94%	57.98%	15.71%
	9이닝 당 홈런 수	13.75%	19.8%	31.1%
	피출루율	49.18%	15.5%	26.88%
	9이닝 당 삼진 수	22.13%	6.71%	26.31%

# 분석 결과 및 투수 평가

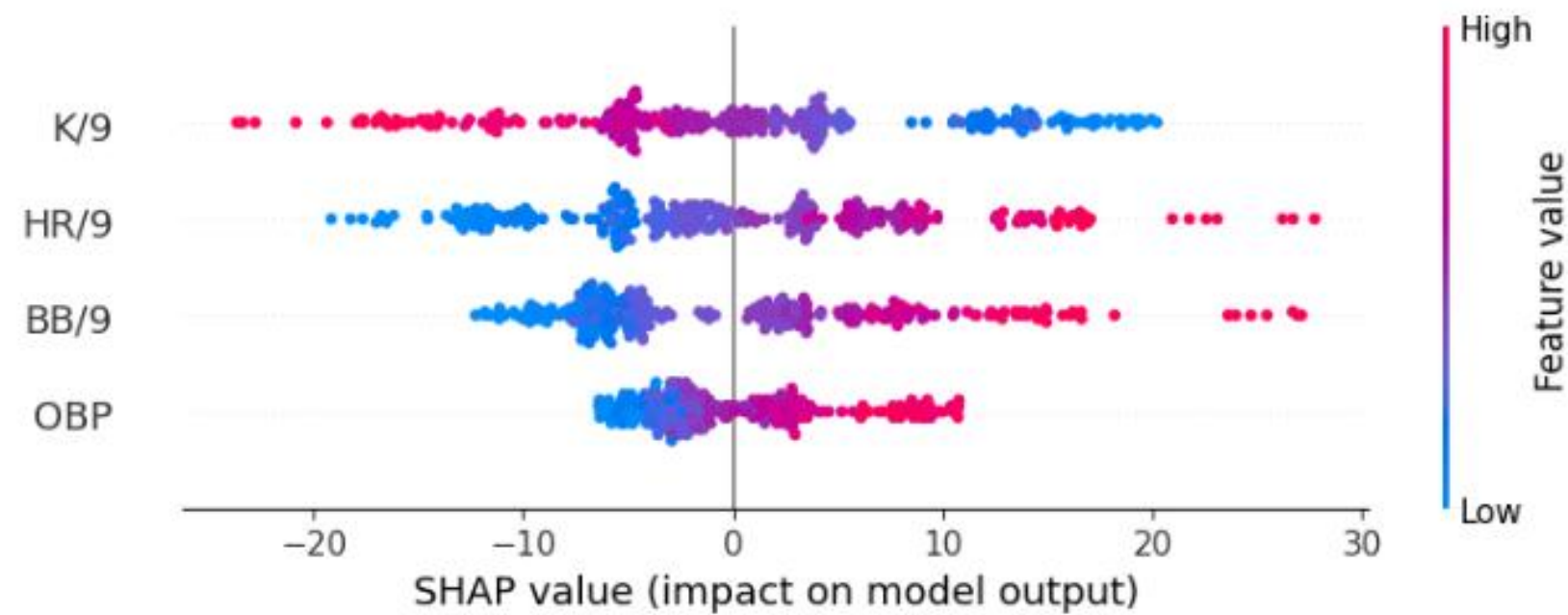
## 변수 영향력 측정(기계학습 모델)

분석대상	변수	XGBoost	Random Forest	LightGBM
구원투수의 조정된 평균자책점	9이닝 당 실점	82.8%	97.81%	58.66%
	피장타율	8.54%	1.26%	25.75%
	잔루율	8.67%	0.09%	15.59%

분석대상	변수	XGBoost	Random Forest	LightGBM
구원투수의 조정된 수비무관 평균자책점	삼진/볼넷 비율	61.26%	56.6%	24.91%
	9이닝 당 홈런 수	10.74%	31.17%	35.53%
	삼진율	11.54%	9.56%	15.75%
	피장타율	9.37%	2%	10.34%
	피출루율	7.09%	0.67%	10.48%

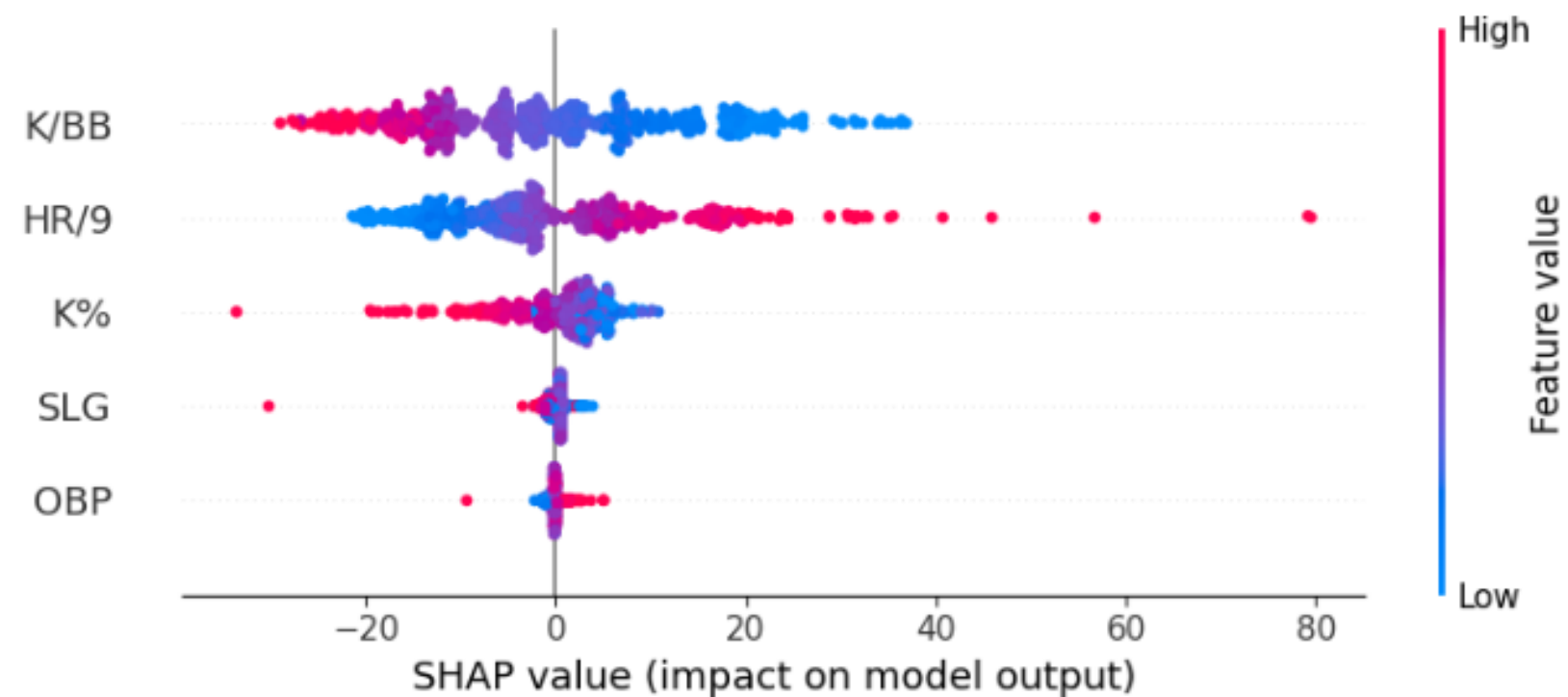
# 분석 결과 및 투수 평가

## 변수 영향력 측정(기계학습 모델)



Ex. 선발투수의 개선된 수비무관 평균자책점 분석 시

- '9이닝 당 삼진 수'은 작을수록 불리
- '9이닝 당 홈런 수', '9이닝 당 볼넷 수', '피출루율'은 클수록 불리



Ex. 구원투수의 개선된 수비무관 평균자책점 분석 시

- '삼진/볼넷 비율', '삼진율'은 작을수록 불리
- '9이닝 당 홈런 수', '피출루율', '피장타율'을 클수록 불리

# 분석 결과 및 투수 평가

## 모델 분석결과 비교

생성지표	Best Model	양의 상관계수	RMSE	결정계수
선발투수 개선된 평균자책점	XGBoost	0.9759	0.2196	0.9518
구원투수 개선된 평균자책점	XGBoost	0.974	0.2282	0.9479
선발투수 개선된 수비무관 평균자책점	XGBoost	0.9739	0.2285	0.9478
구원투수 개선된 수비무관 평균자책점	XGBoost	0.9767	0.2159	0.9534

전체 모델 중 XGBoost 모델에서 생성한 지표가 상관계수, RMSE, 결정계수 값에서 모두 높음

- XGBoost 모델이 가장 좋은 성능의 지표를 생성하였음



# 분석 결과 및 투수 평가

## 지표 비교

연구에서 투수의 개선된 평균자책점&개선된 수비무관 평균자책점 지표를 생성

- 기존의 평균자책점&수비무관 평균자책점과 비교함으로 투수 성적에 대한 평가가 개선되었는지 확인

## 사용 지표

평균자책점(ERA), 수비무관 평균자책점(FIP)

## 평가 방법

양의 상관계수, 결정계수, RMSE를 고려하여 가장 높은 설명력으로 투수 가치를 설명할 수 있는 지표 선택



# 분석 결과 및 투수 평가

## 지표 비교

분석대상	지표	양의 상관계수	RMSE	결정계수
선발투수의 평균자책점과 비교한 지표 평가	ERA	0.9619	0.2759	0.9239
	Our ERA metric	0.9759	0.2196	0.9518
선발투수의 수비무관 평균 자책점과 비교한 지표 평가	FIP	0.9366	0.3561	0.8733
	Our FIP metric	0.9739	0.2285	0.9478

선발투수의 평균자책점과 수비무관 평균자책점을 연구에서 개선한 투수 지표와 비교한 결과

- ‘평균자책점’과 ‘수비무관 평균자책점’과 비교하였을 때, 지표 설명력이 더 향상됨을 보임



# 분석 결과 및 투수 평가

## 지표 비교

분석대상	지표	양의 상관계수	RMSE	결정계수
구원투수의 평균자책점과 비교한 지표 평가	ERA	0.97	0.2448	0.94
	Our ERA metric	0.974	0.2282	0.9479
구원투수의 수비무관 평균자책점과 비교한 지표 평가	FIP	0.9241	0.3895	0.8482
	Our FIP metric	0.9767	0.2159	0.9534

구원투수의 평균자책점과 수비무관 평균자책점을 연구에서 개선한 투수 지표와 비교한 결과

- ‘평균자책점’과 ‘수비무관 평균자책점’과 비교하였을 때, 지표 설명력이 더 향상됨을 보임



# 분석 결과 및 투수 평가

## 선수 평가 예시(선발 투수 A 예시)

시즌	구장 파크팩터	평균자책점	9이닝 당 실점 수	9이닝 당 볼넷 수	피장타율	삼진율	조정 평균자책점	개선된 평균자책점
2022	1.04	4.56	4.87	2.36	0.440	20.2	110.42	92.78
2024	0.97	4.31	4.98	1.19	0.467	14.1	89.26	116.77
변수 중요도(사용 변수만)			76.97%	6.57%	9.09%	7.36%		

선발투수 A의 개선된 평균자책점이 향상된 때는 평균자책점과 9이닝 당 실점 수의 차이가 크지 않았음  
나머지 3개의 변수도 리그 평균보다 좋았던 시즌이었고, 타자친화구장을 사용하여 값이 조정되었음

선발투수 A의 개선된 평균자책점이 하락한 때는 평균자책점과 9이닝 당 실점 수의 차이가 존재하기 시작함  
나머지 3개의 변수 중, 삼진율은 리그평균보다 못 미치는 수치였음  
투수친화구장을 사용하여 값이 조정됨

# 분석 결과 및 투수 평가

## 선수 평가 예시(구원 투수 비교 예시)

이름 및 시즌	구장 파크팩터	평균자책점	9이닝 당 실점 수	9이닝 당 볼넷 수	피장타율	조정 평균자책점	개선된 평균자책점
김00(2023)	0.95	2.18	2.18	3.33	0.306	54.66	42.01
홍00(2024)	0.95	2.73	3.34	5.01	0.359	57.28	71.53
변수 중요도(사용 변수만)			82.8%	8.67%	8.54%		

김00선수는 평균자책점과 9이닝 당 실점 수가 동일하며, 리그 최상위권 수준을 기록하였음  
투수친화구장에서 경기를 하였지만, 나머지 지표도 리그평균보다 좋으며 개선된 투수 지표값이 향상됨

홍00선수는 평균자책점과 9이닝 당 실점 수의 차이가 존재하며, 다른 지표도 김00선수보다 높음  
투수친화구장에서 홈경기를 진행해 식 산정에 불리하게 작용되며, 개선된 투수 지표값은 향상되지 못함

# 분석 결과 및 투수 평가

## 구장 파크팩터 영향 확인

분석대상	결과	구장 파크팩터 > 1.05	1 < 구장 파크팩터 < 1.05	0.95 < 구장 파크팩터 < 1	구장 파크팩터 < 0.95	전체 수
선발투수 개선된 평균자책점	성적 향상	64	66	29	8	167
	성적 감소	5	21	79	36	141
	합산	69	87	108	44	308
구원투수 개선된 평균자책점	성적 향상	85	108	68	17	278
	성적 감소	31	40	129	70	270
	합산	116	148	197	87	548

선발투수의 개선된 평균자책점 지표를 분석 시, 성적이 향상된 167명의 선발투수 중 130명은 타자친화구장에서 경기를 진행  
반대로 성적이 감소된 141명의 선발투수 중 115명은 투수친화구장에서 경기를 진행

구원투수의 개선된 평균자책점 지표를 분석 시, 성적이 향상된 278명의 구원투수 중 193명은 타자친화구장에서 경기를 진행  
반대로 성적이 감소된 270명의 구원투수 중 199명은 투수친화구장에서 경기를 진행



# 분석 결과 및 투수 평가

## 구장 파크팩터 영향 확인

분석대상	결과	구장 파크팩터 > 1.05	1 < 구장 파크팩터 < 1.05	0.95 < 구장 파크팩터 < 1	구장 파크팩터 < 0.95	전체 수
선발투수 개선된 수비무관 평균자책점	성적 향상	62	76	17	5	160
	성적 감소	7	11	91	39	148
	합산	69	87	108	44	308
구원투수 개선된 수비무관 평균자책점	성적 향상	105	129	34	10	278
	성적 감소	11	19	163	77	270
	합산	116	148	197	87	548

선발투수의 개선된 수비무관 평균자책점 지표를 분석 시, 성적이 향상된 160명의 선발투수 중 138명은 타자친화구장에서 경기를 진행

반대로 성적이 감소된 148명의 선발투수 중 130명은 투수친화구장에서 경기를 진행

구원투수의 개선된 수비무관 평균자책점 지표를 분석 시, 성적이 향상된 278명의 구원투수 중 234명은 타자친화구장에서 경기를 진행

반대로 성적이 감소된 270명의 구원투수 중 240명은 투수친화구장에서 경기를 진행



# 결론

## 개선된 투수 지표를 사용하였을 때 장점

- 여러 변수를 고려하여 투수의 객관적 실력을 평가하는 설명력을 향상시킴(다중 변수를 사용해 선수를 평가할 수 있음)
- 구장 파크팩터에 영향을 받는 점을 확인하였음(타자친화적인 구장에서 투구하는 투수들의 평가를 향상)

## 기대효과

선수들의 성적 변화를 파악하는 것이 여러 각도로 이루어지며, 분석이 세분화 될 것을 기대

자유계약(FA)시장에서의 적절한 투수 평가가 이루어져 구단과 선수는 계약하는데 각 걱정선을 둘 수 있음

## 한계 및 보완점

타구 속도, 구장의 세부 특징을 고려하지 못했기 때문에 완전한 지표는 아님 → 기계학습을 이용한 세부 연구가 필요

구장 파크팩터 산정식은 정답이 없으므로, 정석에 근접한 산정식이 필요 → 연구가 지속적으로 필요

MLB리그에는 statcast라는 데이터를 사용하여 더욱 정교한 지표 및 투수평가방법을 고려 중,  
KBO에서도 도입 및 향후 발전하였을 때 더욱 정확한 투수평가방법을 만들 수 있을 것으로 기대



# 참고문헌

- [1]J. Bill. 2010. The new Bill James historical baseball abstract. Simon and Schuster
- [2]한국프로야구에서 타자능력지수 제안-대체선수대비승수 (WAR) 을 중심으로.  
이제영, 김현규. 2016. 한국프로야구에서 타자능력지수 제안-대체선수대비승수 (WAR) 을 중심으로. 응용통계연구, 29(7), 1271-1281.
- [3]주윤태, 류현지, 서상훈. 2023. 기계학습을 이용한 한국 프로야구 신인 투수들의 유형화 및 잠재력 평가. 코칭능력개발지, 25(2), pp 190-197.
- [4]Konaka, E. 2021. Park factor estimation improvement using pairwise comparison method. arXiv preprint arXiv:2109.09287.
- [5]Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 58 Issue 1, p 267-288.
- [6]Vidaurre, D., Bielza, C., Larranaga, P. 2013. A survey of L1 regression. International Statistical Review, Volume 81 Issue 3, p 361-387.
- [7] Zou, H., Hastie, T. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology. Volume 67 Issue 2, p 301-320.
- [8] Breiman, L. 2001. Random forests. Machine learning, Volume 45, p 5-32.
- [9] Chen, T., Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, p. 785-794.
- [10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... and Liu, T. Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems. 30.

# 참고문헌

- [12] Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike (pp. 199-213). New York, NY: Springer New York.
- [13] Acharya, R. A., Ahmed, A. J., D'Amour, A. N., Lu, H., Morris, C. N., Oglevee, B. D., ... & Swift, R. N. (2008). Improving major league baseball park factor estimates. *Journal of Quantitative Analysis in Sports*, 4(2).
- [14] Konaka, E. (2021). Park factor estimation improvement using pairwise comparison method. arXiv preprint arXiv:2109.09287.

# 사이트 및 참고서적

[4]Pitching and Defense: How Much Control Do Hurlers Have?

[11]스카우트 리포트 2018, 레이더는 타구의 질을 어떻게 평가할까?

[15] [baseball-reference.com/about/parkadjust.shtml](http://baseball-reference.com/about/parkadjust.shtml)



# 부록(KBO리그 5년간 팀별 파크팩터)

Team	season	Adjust park factor
LG	20	0.9362
	21	0.9531
	22	0.945
	23	0.9572
	24	0.9735

Team	season	Adjust park factor
KT	20	1.001
	21	1.0023
	22	1.058
	23	1.0741
	24	1.0389



# 부록(KBO리그 5년간 팀별 파크팩터)

Team	season	Adjust park factor
두산	20	0.9321
	21	0.9229
	22	0.9085
	23	0.9494
	24	0.9538

Team	season	Adjust park factor
SSG	20	1.0495
	21	1.0253
	22	1.0285
	23	1.0207
	24	1.052



# 부록(KBO리그 5년간 팀별 파크팩터)

Team	season	Adjust park factor
롯데	20	1.0278
	21	1.0374
	22	1.06
	23	1.0457
	24	1.0561

Team	season	Adjust park factor
한화	20	0.989
	21	0.9372
	22	0.9586
	23	0.9682
	24	0.9853



# 부록(KBO리그 5년간 팀별 파크팩터)

Team	season	Adjust park factor
NC	20	1.053
	21	1.0712
	22	1.0396
	23	1.0191
	24	0.9739

Team	season	Adjust park factor
키움	20	0.9601
	21	0.9823
	22	0.9573
	23	0.9877
	24	0.9463





# 부록(KBO리그 5년간 팀별 파크팩터)

Team	season	Adjust park factor
기아	20	1.016
	21	0.9947
	22	0.9985
	23	0.965
	24	0.9881

Team	season	Adjust park factor
삼성	20	1.0999
	21	1.0818
	22	1.0541
	23	1.029
	24	1.0639



**감사합니다.**

