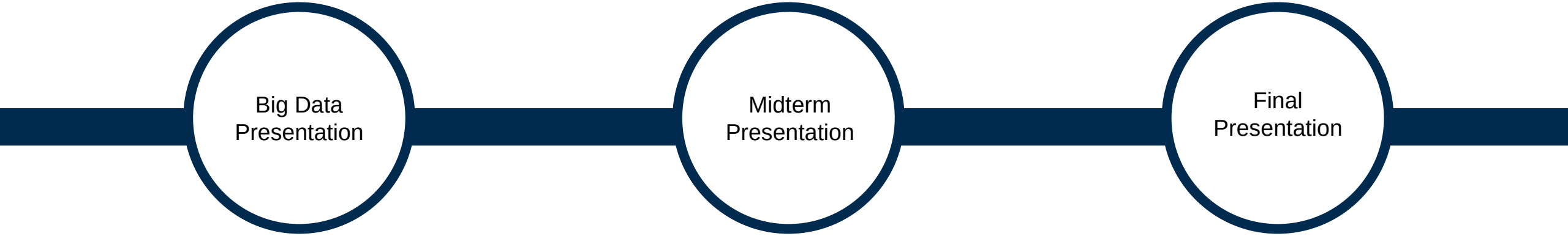
2023 AI·DS Capstone

MLB's on-base and out prediction model by pitcher and batter type using machine learning

안동수 . 김지민 . 손주찬

Research History



Contents

Chapter 1. Introduction

Chapter 2. Model Development

Chapter 3. Conclusion and Application

Chapter 1

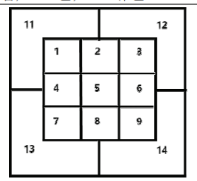
Introduction

Introduction

Limitations of Previous Studies

1. Unable to reflect the complex situation of the game

항목		내용	속성(종류)	타입
예측결과 · 종속	event	▷ 투구결과 ▷ hit-into-play 된 타구 중 안타와 홈런만 측정	▷ 안타 : 1루타(single), 2루타(double), 3루타(triple) ▷ 홈런(homerun)	명목
예측을 위한 투입변수 · 독립	pitch_type	구종	▷ 직구 : 패스트볼(4심), 싱커(2심), 커터, 컷패스트볼 ▷ 변화구 : 슬라이더, 커브, 너클커브, 슬로우커브, 너클, 이퍼스, 슬로브, 스위퍼 ▷ 오프스피드 : 스플릿핑거, 체인지업, 포크볼, 스크류볼	명목
	release_speed	투구속도	단위 : 마일(mph)	연속
	zone	투구존	1번~14번	명목
	stand	타자위치	좌타 / 우타	명목
	p_throws	투수주손	좌완 / 우완	명목
	Balls	볼카운트	0~3	이산
	strikes	스트라이크카운트	0~2	이산



〈투구존〉

Excluding various variables that could affect the outcome of the pitch

+

The situation to predict is fragmented
as it is limited to the hitting situation

⬇

Limitation of not capturing the complexity of baseball,
where a single change in a small variable can lead to a variable of situations

2. Overlooking the Impact Player-Specific Characteristics

			구 질					전체
			직구	커브	슬라이더	체인지업	싱커	
선수	S	빈도	479	241	0	187	5	912
		전체%	31.6%	1.8%	0%	7.1%	0.2%	34.5%
	P	빈도	834	47	725	123	0	1729
		전체%	31.6%	1.8%	27.5%	4.7%	0%	65.5%
전체		빈도	1313	288	725	310	5	2641
		전체%	49.7%	10.9%	27.5%	11.7%	0.2%	100.0%

Various types of pitchers depending on the player's style

+

Matchups may differ based on player types within positions,
influencing strategic approaches

⬇

Previous research has not adequately considered variables
stemming from player characteristics

참고 문헌:
1). 조선미, 김주학, 강지연, 김상균. "머신러닝(XGBoost)기반 미국프로야구(MLB)의 투구별 안타 및 홈런 예측 모델 개발." 한국체육측정평가학회지, vol. 25, no. 1, 2023, pp. 65-76.
2). 손혁. "프로야구 투수유형과 구질과의 관계." 국내석사학위논문 고려대학교 교육대학원, 2004. 서울
3). 황수웅. 불확실성(uncertainty)을 고려한 스포츠 빅데이터 분석: Bayesian 추정과 Deep Learning을 활용한 프로야구 심판의 Ball/Strike 판정 평가 모델 개발. 서울대학교 대학원. 2023

Introduction

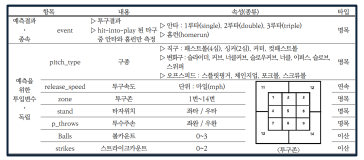
Research Novelty

Diversification of Input Variables

Predictions on multiple Scenarios

Factors influencing outcomes for different player types

기존 변수



+

신규 변수

공의 상하 움직임
투구 폼
투수 익스텐션
투수 릴리스 포인트
주자 여부
점수, 이닝
아웃카운트

Consider the factors that can affect pitching results in more ways than previous

기존 연구

타격 결과 (안타, 홈런)
투구 결과 (스트라이크, 볼)

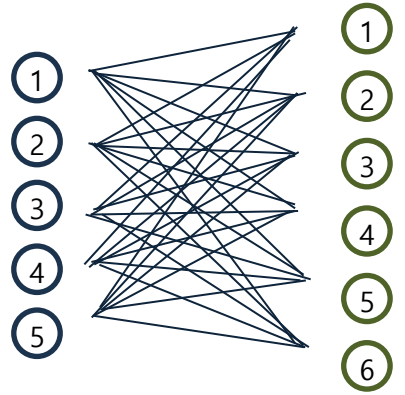
not a comparable result
Predict the number of cases in the same situation

본 연구

출루
(안타, 홈런, 볼넷)
+
아웃
(삼진, 뜬공, 병살...)

Predictions including various situations that may occur in baseball

타자 유형 **투수 유형**



Considering the characteristics of each player's type and the variables of the characteristics caused by this type

Introduction

Research Importance

Importance of identifying on-base/out indicators

Base Runners			2010-2015			1993-2009			1969-1992			1950-1968		
1B	2B	3B	0 outs	1 outs	2 outs	0 outs	1 outs	2 outs	0 outs	1 outs	2 outs	0 outs	1 outs	2 outs
__	__	__	0.481	0.254	0.098	0.547	0.293	0.113	0.477	0.252	0.094	0.476	0.256	0.098
1B	__	__	0.859	0.509	0.224	0.944	0.565	0.245	0.853	0.504	0.216	0.837	0.507	0.216
__	2B	__	1.100	0.664	0.319	1.175	0.723	0.349	1.102	0.678	0.325	1.094	0.680	0.330
1B	2B	__	1.437	0.884	0.429	1.562	0.966	0.471	1.476	0.902	0.435	1.472	0.927	0.441
__	__	3B	1.350	0.950	0.353	1.442	0.991	0.388	1.340	0.943	0.373	1.342	0.926	0.378
1B	__	3B	1.784	1.130	0.478	1.854	1.216	0.533	1.715	1.149	0.484	1.696	1.151	0.504
__	2B	3B	1.964	1.376	0.580	2.053	1.449	0.626	1.967	1.380	0.594	1.977	1.385	0.620
1B	2B	3B	2.292	1.541	0.752	2.390	1.635	0.815	2.343	1.545	0.752	2.315	1.540	0.747

□□: http://www.tangotiger.net/re24.html

- 해당 주루/아웃 상태에서 해당 이닝이 끝날 때까지 득점할 확률

Differences in scoring probability depending on the presence or absence of runners in the same out situation

→ Identify the indicators of out and on-base = Factors that directly lead to victory or defeat

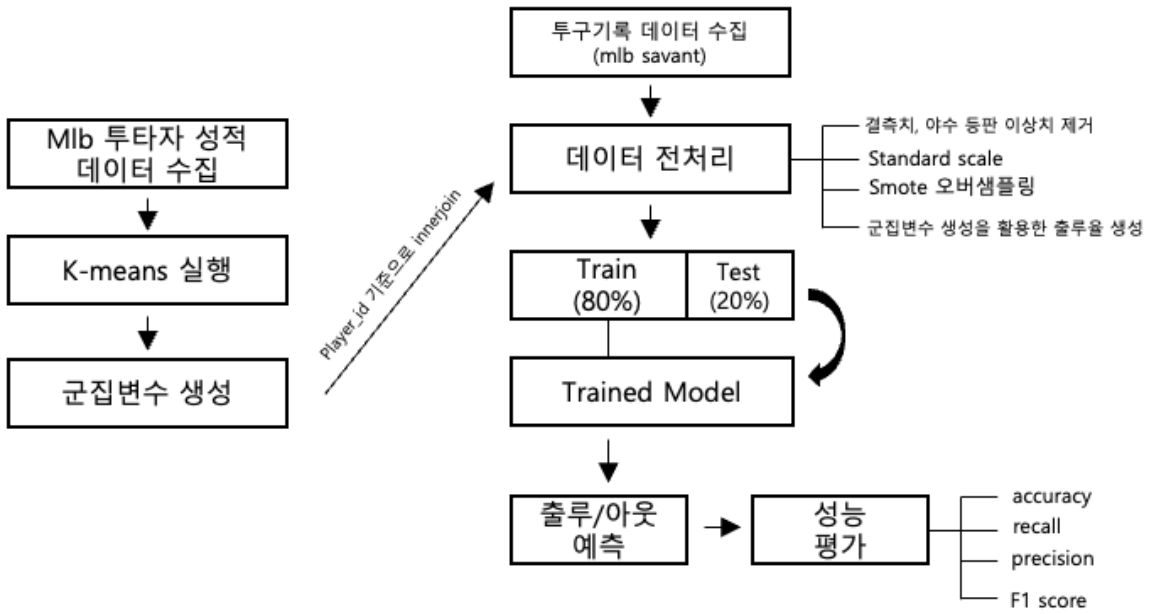
Create a strategy based on data

□□: https://www.gqkorea.co.kr/2020/03/05/

If knowing the type and type of player, rational decisions can be made in establishing strategies such as batter placement and pitcher replacement strategies

Introduction

Model development plan



Step 1. Data Collection

Step 2. Data Preprocessing

Step 3. Creation of New Independent Variables using K-means

Step 4. Data Integration

Step 5. Variable Scaling & Oversampling

Step 6. On-Base Percentage Variable Generation

Step 7. Train Set & Test Set Creation

Step 8. Model Training

Chapter 2.

Model Development

Model Development

Step 1. Data Collection

Collection of pitch-related data from Major League Baseball's official record website "MLB Savant."

Search Results

Rk.	Player	Pitches	Total	Pitch %	
1	Cole, Gerrit RHP	3281	3281	100.0	Graphs
2	Cease, Dylan RHP	3262	3262	100.0	Graphs
3	Gallen, Zac RHP	3248	3248	100.0	Graphs
4	Castillo, Luis RHP	3207	3207	100.0	Graphs
5	Mikolas, Miles RHP	3197	3197	100.0	Graphs
6	Giolito, Lucas RHP	3190	3190	100.0	Graphs
7	Webb, Logan RHP	3182	3182	100.0	Graphs
8	Snell, Blake LHP	3168	3168	100.0	Graphs
9	Lynn, Lance RHP	3167	3167	100.0	Graphs
10	Wheeler, Zack RHP	3155	3155	100.0	Graphs
11	Bassitt, Chris RHP	3139	3139	100.0	Graphs
12	Keller, Mitch RHP	3119	3119	100.0	Graphs
13	Strider, Spencer RHP	3100	3100	100.0	Graphs
14	Nola, Aaron RHP	3087	3087	100.0	Graphs
15	Burnes, Corbin RHP	3079	3079	100.0	Graphs

출처: https://baseballsavant.mlb.com/

List of data collected (2020-2022) → 421,879 rows

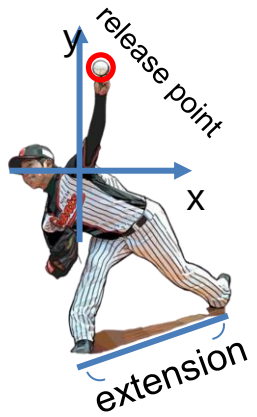
1. Dependent Variables:

- On-Base / Out

2. Independent Variables:

- Strike, Ball
- Pitch Velocity, Pitch Type
- Vertical and Horizontal Ball Movement
- Pitching Form, Pitcher Extension, Pitch Release Point
- Game Situation at the Time of Pitching

(Presence of Runners, Game Score, Inning, Out Count)



참고문헌:
SmartPitch: Applied Machine Learning for Professional Baseball Pitching Strategy. (Otremba Jr, 2022)

Model Development

Step 2. Data Preprocessing

1). Removing Outliers

predict the on-base between normal positions (pitcher vs. hitter)

→ Excluding cases where a batter takes the mound as a pitcher as an outlier

- Criteria for judging the batter's mound.

- MLB Pitcher Lowest Pitcher Under 69.9 Mile In 2022

- 2021 MLB Pitcher Lowest Pitcher Under 64.7 Mile

- 2020 MLB Pitcher Lowest Pitcher Less Than 65 Mile

- The above balls are judged to have been thrown by the other

→ Delete pitches below the minimum pitcher's ball speed by year as outliers

2). Removing Missing Value

The percentage of missing values in all variables is 0.3% or less

→ Decide that there will be no impact on the model, Remove all missing values

	결측치 비율 (%)
release_pos_x	0.013823
release_pos_z	0.013823
pfx_x	0.001728
pfx_z	0.000494
release_spin_rate	0.328060
release_extension	0.140209
release_pos_y	0.013823
spin_axis	0.328060

3). Preprocessing categorical data

Mapping to number the categorical data

- Pitcher Handedness: Right-handed/Left-handed → 1 / 2
- Batter Handedness: Right-handed/Left-handed → 1 / 2
- Pitch Type: Fastball → 1, Offspeed → 2, Breaking Ball → 3
- On-Base: Hit, Home Run, Walk, Hit by Pitch → 1
- Out: Double Play, Triple Play, Flyout, Groundout → 0

Pitch_type	Release_speed	On_1b	...	release_pos_x	release_pos_y	events
3	90.0	0	...	-2.20	6.49	1
1	95.2	0	...	-2.33	6.41	0
1	92.2	1	...	-2.13	6.48	0
3	83.7	0	...	-2.28	6.33	0

독립 변수 종속 변수

Model Development

Step 3. Generate new variables using K-means clustering on a per-player basis

1). Grouping by players' performance

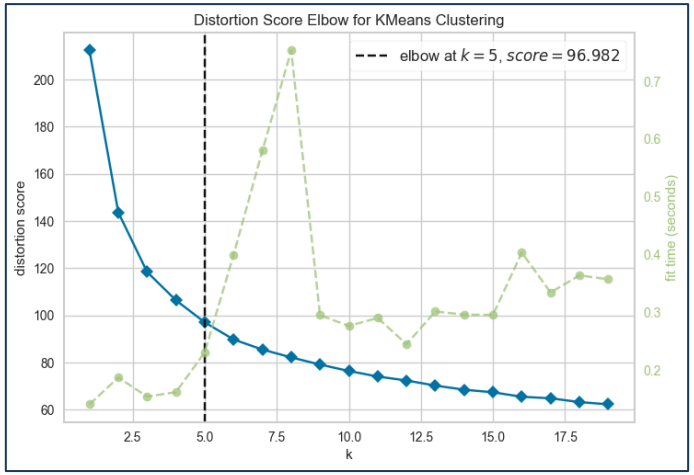
1. Using average performance data for pitchers and batters for 2020 to 2022

- Batters: strikeout rate, walk on base rate, home run, ops, quality of batting
- Pitcher: Fastball and breaking ball restraint, hit rate, ERA

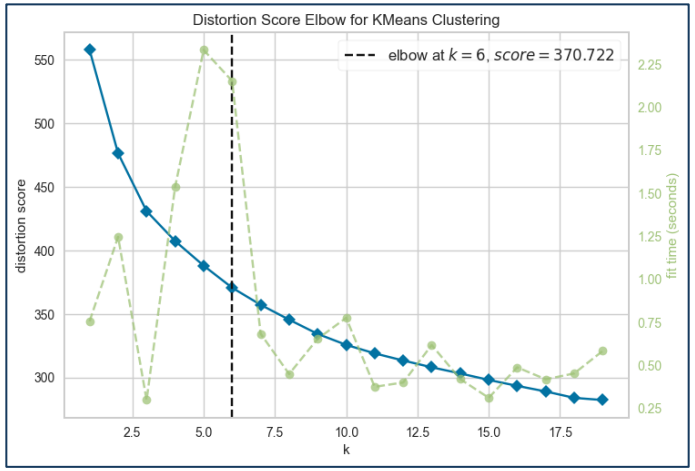
2.. Selection of the number of clusters using the "Elbow Method"

- 5 batters
- 6 pitchers

3. Clustering



Selection of the number of batter clusters



Selection of the number of Pitcher Clusters

참고문헌:
1). 한국프로야구에서 타자능력의 측정 (이장택, 2014)
2). Investigating Major League Baseball Pitchers and Quality of Contact through Cluster Analysis"(Marcou,2020)

Model Development

Step 3. Generate new variables using K-means clustering on a per-player basis

2). Check the characteristic of cluster

- characteristic of batter cluster

군집	키워드	대표 선수	홈런	삼진율	볼넷율	타율	출루율	장타율	OPS	Babip	Pull Percent	Line drive percent
1번	당겨치기, 낮은 타율, 높은 볼넷 출루율	최지만	7.06	27.9%	9.1%	0.220	0.300	0.398	0.698	0.276	42.49%	23.2%
2번	전반적 낮은 수치	쓰쓰고 요시토모	2.46	26.3%	7.5%	0.195	0.264	0.291	0.556	0.255	37.22%	22.0%
3번	낮은 삼진율 높은 라인드라이브 타율	김하성	6.66	20.3%	7.4%	0.255	0.318	0.386	0.704	0.306	38.71%	25.0%
4번	높은 타율, 타점, 홈런 핵심 타자	오타니 쇼헤이	18.6	22.3%	9.6%	0.258	0.337	0.467	0.804	0.297	39.29%	23.98%
5번	메이저리그 적응 실패, 매우 낮은 성적	린즈웨이	0.148	45.27%	3.4%	0.114	0.149	0.146	I 0.296	0.235	23.82%	15.7%

- characteristic of pitcher cluster

군집	키워드	대표 선수	직구 구속	변화구 구속	삼진율	볼넷 허용률	방어율	허용타구 속도	포심 구사율	싱커 구사율	슬라이더 구사율	직구 구사율	땅볼유도
1번	싱커 구사, 땅볼 유도형	펠릭스 페나	93	83	20%	8.9%	4.36	88.2	12.6	45.8	24.5	59.3	51.4
2번	빠른 포심, 탈삼진 능력	맥스 슈어저	94.3	84.2	24.6%	9.6%	4.15	88.8	51.3	9.1	22.6	56.6	39.9
3번	낮은 구속, 정교한 제구력	류현진	90.3	78.8	21.1%	8.2%	4.38	88.1	26.2	20.9	9.9	69.8	42.8
4번	메이저리그 경력 적지만 방어율 높음	조던 야마모토	91.9	81.3	18%	9.9%	7.51	90.3	43.4	14.1	21.5	56.7	36.3
5번	마무리 투수, 높은 삼진율	에드윈 디아즈	93.7	84.2	25.5%	9.5%	3.95	88.3	34.6	13.1	46.5	43.7	43.2
6번	다양한 구종, 선발급 능력	오타니 쇼헤이	92.9	80.4	24.4%	9.1%	4.13	88.5	45.1	11.6	14.8	55.5	41.7

Model Development

Step 3. Generate new variables using K-means clustering on a per-player basis

2). Check the characteristic of cluster

- Characteristics by Batter Cluster

선구안이 좋은 타자그룹

- 낮은 타율(0.220)
- 높은 볼넷율(9.1%)
- 높은 출루율(0.300)

볼을 잘 골라내는 능력을 가졌고 출루율이 높은 타자 그룹
Ex). 최지만 선수

컨택트형 타자 그룹

- 평균 이상의 타율(0.255)
- 낮은 삼진율 (20.3%)
- 높은 라인드라이브 타구율 (25%)

볼 컨택이 좋기 때문에 좋은 타구질을 만들어내는 타자
Ex) 김하성 선수 류

홈런형 파워타자 그룹

- 높은 타율 (0.258)
- 높은 타점 (58)
- 높은 홈런 수 (59)

팀내 중심타선을 담당할 정도의 파워를 가진 4번타자형 타입
Ex) 오타니 쇼헤이 선수

Model Development

Step 3. Generate new variables using K-means clustering on a per-player basis

2). Check the characteristic of cluster

- Characteristics by Pitcher Cluster

땅볼 유도형 투수그룹

- 메이저리그 평균 구속(93마일)
- 땅볼 유도 구종인 싱커의 높은 구사율 (45.8%)
- 높은 땅볼유도 비율 (51.4%)

→ 싱커를 주로 구사하는 땅볼유도형

Ex) "땅볼 유도형" 펠릭스 페냐 선수

직구 위주 삼진형 투수그룹

- 높은 포심 구사율 (51%)
- 높은 삼진율 (24.6%)
- 가장 빠른 직구구속(94.3마일)

→ 직구 구위로 승부하는 투수

Ex) "메이저리그 최고의 포심으로 선정" 맥스 슈어저

제구형 투수그룹

- 낮은 직구속도(90마일)
- 낮은 변화구 속도(78마일)
- 낮은 볼넷 허용률(8.2%)

→ 변화구와 직구 사이에 큰 속도편차를 활용하고 정교한 제구력으로 승부

Ex) "칼제구 " 류현진 선수

직구, 슬라이더 등 구위가 좋은 삼진형 투수그룹

- 적은 평균 이닝 소화(39.5이닝)
- 가장 높은 삼진율 (25.5%),
- 가장 낮은 방어율 (3.95),
- 낮은 빠른 타구 허용률 (36.8%)
- 직구 구사율 (43%)
- 슬라이더 구사율(46.5%)

→ 직구, 슬라이더를 주로 구사하는 마무리 투수 타입

Ex) "절대마무리" 에드윈 디아즈

다양한 구종을 구사하는 에이스 선발투수그룹

- 낮은 방어율(4.13)
- 많은 평균이닝(63이닝)
- 삼진율 24%
- 모든 구종 구사비율 최소 10%이상

→ 다양한 구종을 던지는 에이스 선발투수

Ex) "MVP" 오타니 쇼헤이

Model Development

Step 3. Generate new variables using K-means clustering on a per-player basis

3). Create cluster variables

Player_id	p_formatted_ip	k_percent	bb_percent	...	Pitcher_cluster_label
424144	18.0	19.4	8.3	...	0
425794	154.1	19.8	6.1	...	2
425844	125.0	18.0	4.3	...	5
429722	65.1	18.8	7.	...	4

Identify the player's unique number for each cluster and assign a variable

Model Development

Step 4. Data merging

Pitch_type	Release_speed	On_1b	...	Batter	Pitcher
2	84.6	1	...	542303	601713
2	84.7	1	...	518692	601713
1	91.7	1	...	645277	601713
1	90.6	0	...	542225	601713

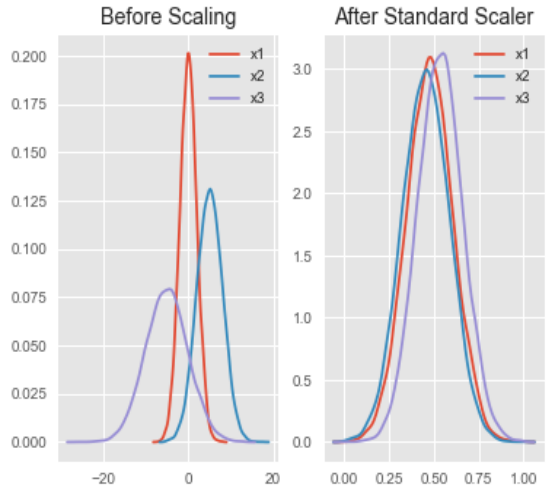
Player_id	p_formatted_ip	k_percent	bb_percent	...	Pitcher_cluster_label
424144	18.0	19.4	8.3	...	0
425794	154.1	19.8	6.1	...	2
425844	125.0	18.0	4.3	...	5
429722	65.1	18.8	7.	...	4

Add a cluster number corresponding to the unique identifiers of the Batter and Pitcher variables in the original data and matching rows in the performance data's Player_id.

Step 5. Data Scaling

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean
σ = Standard Deviation



Different variables have different data scales, requiring adjustment for more accurate predictions

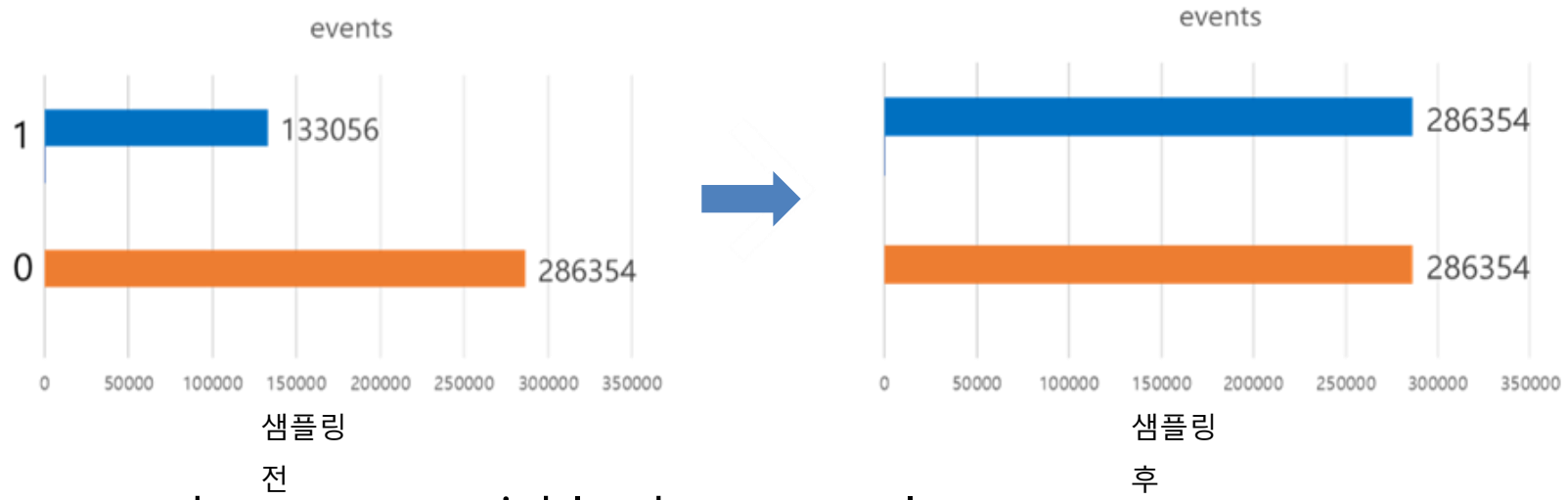
⬇

Using the standard scaler of the scikit-learn library to scale properties to zero and variance to one

Model Development

Step 6. Smote Oversampling

- Using the Smote Oversampling technique to solve the unbalance issues in dependent variables



Step 7. Create on-base rate variables between player types

- Calculate the ratio of on-base and out through the results of each cluster of pitchers and batters

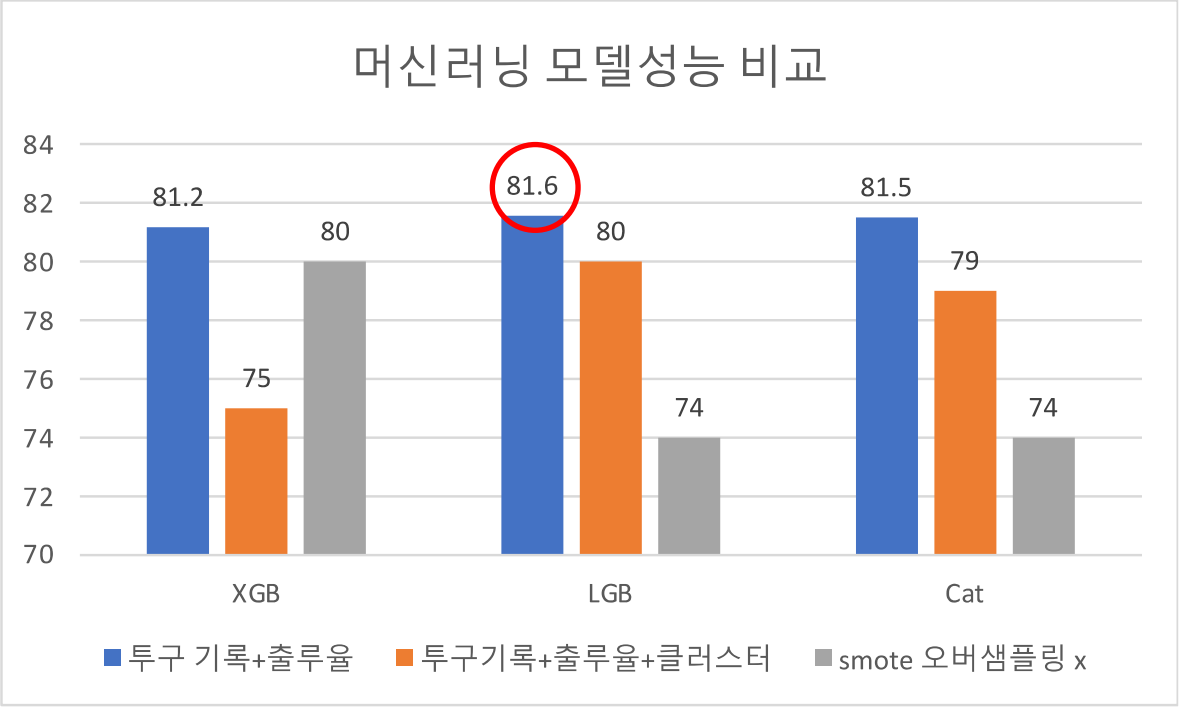
→ Create on-base percentage calculation results in a new column called 'On_base_ratio' and use them as variables for prediction

- On-base percentage = (on-base)/(number of showdowns between pitcher and batter)
- → Ex) P1 (pitcher cluster 1) vs B1 (baseball cluster 1) = 100 (on-base)/100 (on-base) + 300 (out) = 0.25

Model Development

Step 8. Model Training

항목		내용	
주요 모델 파라미터	n_estimators	300	트리개수
	num_leaves	65	트리의 리프 노드 개수
	max_depth	13	트리의 최대 깊이
	learning_rate	0.1	학습률
	colsample_bytree	0.5	트리별 특성 샘플링 비율
	reg_alpha	0	L1 정규화 파라미터
	reg_lambda	0	L2 정규화 파라미터
학습 태스크	subsample	0.5	전체 데이터 사용
	학습비	8:2	모델 학습과 테스트



- No overfitting due to oversampling

- Adjust hyperparameters to randomsearch cv

참고 문헌:
1). Football Matches Outcomes Prediction Based on Gradient Boosting Algorithms and Football Rating System(Razali, 2021)
2). 머신러닝을 활용한 빅데이터 분석을 통해 KBO 타자의 OPS 예측(한정섭, 2022)

Chapter 3.

Conclusion and Application

Conclusion and Application

1. Identify team resources, recruit players, plan strategies.

- **Understand player types and their compatibility**

1. Collecting power information on what types of players are in the team
2. Understand what types of players your team is struggling against with its current strength.

→ Player type information helps guide team resources and deployment strategies

2023 MLB 신인 투수 센가 코다이

- **Use data to identify rookie pitcher types**

✓ Analyzing 2023 MLB Rookie Pitching with the results

- Senga Kodai → Pitcher cluster 5(Power pitcher type)
- Analyze the characteristics of pitchers in cluster 5 (e.g., Edwin Diaz) to prepare for a rookie matchup.

→ Understand the types of rookies and prepare for the matchup

Conclusion and Application

2. On-base percentage by player type

1. 땅볼 유도형 투수 vs 파워형 타자 :

상대적으로 땅볼 투수형은 컨택트형 타자에 강하고, 파워형 타자 그룹에 약한 면모

(컨택트형 타자에게 출루 허용 → 0.409 / 파워형 타자에게 출루 허용 → 0.458)

"땅볼 유도형"
펠릭스 페냐

VS

"홈런형 파워타자"
오타니 쇼헤이

2. 마무리 투수 vs 파워형 타자 :

상대적으로 마무리 투수형은 컨택트형 타자에 강하고 파워형 타자 그룹에 약한 면모

(컨택트형 타자에게 출루 허용 → 0.413 / 파워형 타자에게 출루 허용 → 0.485)

"마무리 투수"
에드윈 디아즈

VS

"홈런형 파워타자"
오타니 쇼헤이

3. 다양한 구종 구사 가능한 투수

- 선구안형 타자에게 강한 면모: 0.362
- 홈런형 파워 타자에게 강한 면모: 0.372
- 컨택트형 타자에 약한 면모: 0.415

다양한 구종 구사 투수
오타니 쇼헤이

VS

"컨택트형 타자" 김하성

"선구안형 타자" 최지만

"홈런형 파워타자"
오타니 쇼헤이

Conclusion and Application

3. Player-specific analytics

- Clayton Kershaw

- Input : 2020-2022 Kershaw's pitching data + cluster

- Leadoff hitter type: 0.232

- Contact hitter type: 0.267

- Slugger type: 0.157

- Analysis insights

→ Strong against sluggers, but relatively weak against contact batters

→ The need to have a pitching strategy to deal with contact hitters

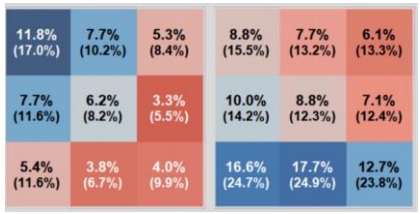
Conclusion and Application

4. Match lineup preparation strategies

- **Utilization Plan**

1. Select an opponent's projected lineup
2. Enter starting pitcher pitching information and pitcher type and batter type
3. Predicting hit results
4. Prepare a preliminary strategy for each opponent

→ Prepare pitch locations and pitch types for each batter with models



Conclusion and Application

5. References for new research

Presents research using a new methodology to predict on-base percentage with pitching history, pitcher and batter clusters

→ Could become a new reference point for sports research

References

1. 김혁주. "한국 프로야구에서 출루 능력과 장타력이 득점 생산성에 미치는 영향," 한국데이터정보과학회지, vol. 23, no. 6, pp. 1165-1174, 2012.

2. 박태신, 김재윤. "머신러닝을 활용한 KBO 외국인 투수 재계약 예측 모형," 한국데이터정보과학회지, vol. 33, no. 6, pp. 963-976, 2022, doi: 10.7465/jkdi.2022.33.6.963.

3. 손혁. "프로야구 투수유형과 구질과의 관계," 국내석사학위논문 고려대학교 교육대학원, 2004. 서울

4. 이장택. "한국프로야구에서 타자능력의 측정," 한국데이터정보과학회지, 25(2), 349-356, 2014.

5. 이승훈, 최형준. "미국 프로야구(MLB) 풀카운트 상황에서 투수의 구질, 구속 변화에 따른 투구 결과 분석," 한국체육과학회지, vol. 28, no. 3, pp. 973-981, 2019, doi: 10.35159/kjss.2019.06.28.3.973.

6. 조선미, 김주학, 강지연, 김상균. "머신러닝(XGBoost)기반 미국프로야구(MLB)의 투구별 안타 및 홈런 예측 모델 개발," 한국체육측정평가학회지, vol. 25, no. 1, 2023, pp. 65-76.

7. 조형석. "MLB 타자들의 스윙존에 따른 스윙선택 성향 분석," 국내석사학위논문 명지대학교 기록정보과학전문대학원, 2021.

8. 최영환. "4차 산업혁명형 ICT기술이 스포츠 분야에 미치는 기술・문화적 동향분석," 한국스포츠학회 16, no.3 (2018): 1-12.

9. 황수웅. "불확실성(uncertainty)을 고려한 스포츠 빅데이터 분석: Bayesian 추정과 Deep Learning을 활용한 프로야구 심판의 Ball/Strike 판정 평가 모델 개발," 서울대학교 대학원, 2023.

10. Albert, Jim. 'Beyond Runs Expectancy'. 1 Jan. 2015: 3 – 18.

11. Marcou, Charlie. "Investigating Major League Baseball Pitchers and Quality of Contact through Cluster Analysis" (2020). Honors Projects. 765.

12. Nathan, Alan M. "What new technologies are teaching us about the game of baseball." Proceedings of the Euromech Physics of Sports Conference. 2012.

13. Otremba Jr., Stephen Eugen. "SmartPitch: Applied Machine Learning for Professional Baseball Pitching Strategy." Massachusetts Institute of Technology, degree of Master of Engineering in Electrical Engineering and Computer Science, 2022. Available at https://hdl.handle.net/1721.1/145144.

THANK YOU