한국체육과학회지, 2024, 제33권 제1호, pp.655-665
The Korea Journal of Sports Science
2024, Vol. 33, No. 1 pp.655-665
https://doi.org/10.35159/kjss.2024.2.33.1.655

머신러닝을 활용한 미국 프로야구의 투수 및 타자의 유형별 출루 및 아웃 예측 모델

안동수[1] · 김지민[2] · 손주찬[3] · 김경외[4*]

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

Ahn, Dong-Su[1] · Kim, Ji-Min[2] · Son, Ju-Chan[3] · Kim, Keung-Oul[4*]

Abstract

The study introduces a novel approach for predicting on-base and out outcomes, employing a combination of pitch data analysis, clustering techniques, and machine learning algorithms. MLB StatCast data from 2020 to 2022 was utilized to analyze and categorize pitcher-batter matchup types based on pitcher-specific pitching characteristics, speed, and player type information, and incorporated as variables in the predictive model. Three machine learning algorithms, namely LightGBM, CatBoost and XGBoost were used to forecast real-time pitch results. This model holds the potential to offer valuable insights for optimizing player deployment, devising game strategies, and enhancing overall baseball analytics.

Key words: Machine learning, MLB, Prediction, On-base and out, Pitch data analysis, Clustering techniques

* awekim@handong.edu
1. 한동대학교(Handong Global University)/학생
2. 한동대학교(Handong Global University)/학생
3. 한동대학교(Handong Global University)/학생
4. 한동대학교(Handong Global University)/교수

Ⅰ. 서 론

전산 기술의 급속한 발전과 더불어 스포츠 분야에서의 데이터 수집 및 분석의 중요성이 점점 더 강조되고 있다(최영환, 2018). 이러한 흐름에 맞추어 야구에서도 단순한 경기 결과를 넘어선 공과 선수의 움직임까지 추적하고 데이터를 수집하는 시대가 도래하였다. 대표적인 예로, 트랙맨 시스템이 있는데, 이 시스템은 레이더 추적을 통해 공의 회전량, 투수가 공을 던지는 지점, 타자의 스윙존 등 보다 세밀한 데이터를 측정한다(이승훈, 2019). 이처럼 야구 전산 기술을 통해 기존 방식보다 더 구체적인 선수 분석과 전략 수립이 가능해졌으며, 이러한 변화는 야구 경기에서 승리에 직접적인 영향을 미치는 요소로 주목받고 있다(최영환, 2018).

그중에서도 프로야구 경기에서의 타격 예측에 관한 연구는 최근 머신러닝과 딥러닝을 활용한 활발한 연구로 주목받고 있다. 몇 가지 주목할 만한 연구들로 그 현황을 파악할 수 있었다. 김민택(2019)과 신동(2022)은 선수의 나이, 타석 수, 안타 수 등 야구 연구에서 전통적으로 사용되는 기초 통계적 요소들을 데이터로 활용하여 타자의 출루율과 장타율의 합인 OPS를 예측하였다. 또한, 조선미(2023)는 XGBoost 기술을 사용하여 투구 속성과 상황 조건에 기반한 투구 기록을 바탕으로 안타와 홈런을 예측하는 모델을 제안하였다.

이처럼 기계학습을 이용한 야구 관련 예측 연구는 활발하게 이루어지고 있지만 아직 연구마다 그 한계점이 명확하게 드러난다. 이들의 공통된 한계를 살펴보자면 야구의 복합적 상황을 충분히 반영하지 못했다는 점이 두드러진다. 야구의 복잡성을 포착하기 위해서는 경기 중에 일어나는 작은 변화 하나라도 감지할 수 있어야 한다. 문형우(2016)의 연구에 따르면 비슷한 상황이라도 주자가 얼마나 진출해 있는지에 따라 기대 득점이나 승리 확률이 확연하게 차이 난다. 김혁주(2012)는 출루 능력과 장타력을 결합한 지표인 'OPS'가 득점 생산성에 미치는 연관성을 비교한 연구를 진행했는데 출루 능력이 다른 지표들보다 팀 득점과 관련이 높다는 것을 확인할 수 있었다. 이러한 근거는 주자 진출 맥락을 이해하는 것이 중요하다는 본연구의 논거를 뒷받침해준다. 그러나 지금까지의 연구들은 야구 경기에서 선수의 기록만 활용하는 단편적인 정보만을 사용했기 때문에 출루해 있는 주자가 한 명이라도 차이가 난다면 승패가 갈리는 야구의 복잡성을 충분히 고려할 수 없었다. 게다가 예측을 목표로 한 상황이 타격 상황 안에서 안타 및 홈런을 예측하거나(조선미, 2023), 투구 결과가 스트라이크인지 볼인지 판단하는 양상을 보였는데 이는 여러 가지 상황이 발생할 수 있는 복합적인 야구의 성격을 놓칠 수 있다는 점에서 보완되어야 한다(황수웅, 2023).

또한 선수들의 개별적 특성에 따른 영향도 간과되었다. 투수의 스타일이나 포지션별 선수 유형과 이에 따른 상성의 효과가 연구에서 빠져 있기에 선수의 특성에서 발생하는 변수에 대한 충분한 고려가 이루어지지 않았다. 타자의 특성별로 스윙 성향을 분석한 조형석(2021)의 연구에 따르면 타자들의 스윙 성향은 상대하는 투수가 좌완인지 우완인지에 따라 개인별 차이가 나타났으며 선수별로 강점을 가지는 스트라이크 존이 있는 것으로 나타났다. 또한 프로야구 투수유형과 구질의 관계를 연구한 손혁(2004)의 연구에서는 투수 별로 구사하는 변화구와 속도 완급 조절 성격이 다르기에 각 투수 별 대응법을 마련할 것을 강조한다. 이처럼 기존 연구들은 다양한 요소들을 반영하지 못함으로 예측 결과에 영향을 줄 수 있는 핵심 요소를 배제했다는 점에서 한계를 보인다.

이에 본 연구에서는 타자의 출루 여부를 예측하기 위한 머신러닝 기반의 새로운 접근 방식을 제안하고자 한다. 먼저 데이터적인 측면에서는 기존의 투구 및 상황 기록 뿐만 아니라 Otremba(2022)가 제안한 연구에서 투수가 타자의 출루를 저지하기 위해 필요하다고 말한 구종, 볼 스피드, 스핀, 투구 위치와 더불어 투수의 투구 자세, 마운드에서의 팔 뻗는 정도, 공의 움직임 등 상세한 투구 추적 데이터를 활용하여 출루 여부와 관련된 상세 정보를 고려한다. 또한 방법론적인 측면에서는 경기 상황의 복잡성과 선수 개인의 특성을 반영하기 위해 클러스터링 분석과 머신러닝 예측 기법을 접목시킨 새로운 접근 방식을 제안한다. 이러한 접근은 야구 데이터 분석의 새로운 관점을 제시함으로써 야구에 대한 이해를 깊이 있게 확장 시키는 데에도 도움을 줄 수 있다. 또한, 이러한 연구를 통해 야구 경기의 승패에 영향을 미치는 요소들을 파악하고, 팀의 전략 수립에 활용함으로써 더 많은 승리를 이끌어 낼 수 있다. 아래 〈그림 1〉은 전체적인 모델의 개발 흐름을 정리한 그림이다.

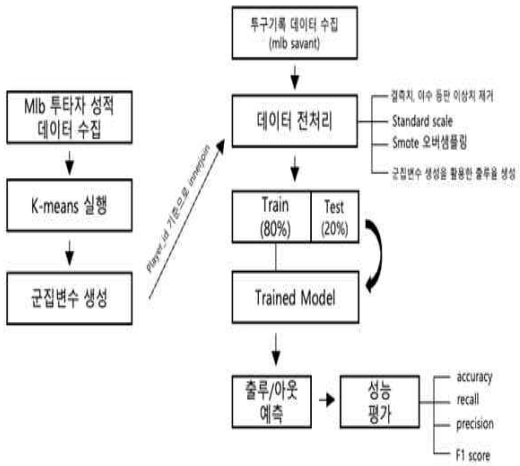

그림 1. 모델 위크 플로우

II. 연구방법

1. 자료수집

본 연구에서는 메이저리그 공식 기록 사이트인 'MLB Stat Cast'에서 제공하는 2020년부터 2022년까지의 전체 야구 기록 데이터를 수집하여 사용하였다. 예측 변수인 출루 더미는 출루(1루타, 2루타, 3루타, 홈런, 볼넷 등)와 아웃(더블아웃, 싱글아웃, 트리플아웃 등) 정보를 기반으로 생성되었으며, 특성 변수로는 좌투/우투, 투수의 릴리즈 포인트, 투수 익스텐션, 공의 무브먼트, 구종, 투구존, 타자 위치, 주자 상황, 볼카운트 등과 같은 투구에 관련된 변수들이 사용되었다.

2. 자료전처리

본 데이터 분석에 앞서 미국 프로야구의 맥락적 정보와 데이터 특성을 활용하여 전처리를 수행하였다. 먼저 데이터 내 존재하는 결측치를 파악한 결과, 결측치가 존재하는 변수들의 결측치 비율이 모두 0.003% 이하인 것으로 나타나 이를 모두 제거하였다. 또한 선수 포지션의 변동이 있는 경우는 분석에서 제외하였다. 프로 야구의 정규 시즌의 경우 많은 경기를 제한된 선수들로 치러야 하기 때문에 타자가 투수로 등판하는 경우가 발생하는데, 이 경우가 많지 않고 일종의 이상치로 볼 수 있기 때문이다. 이를 판단하기 위해 매년 투수의 최저 구속(2022:69.9 마일, 2021: 64마일, 2020: 65마일)보다 구속이 낮은 경우를 타자가 투수로 등판한 경우로 간주하고 제거하였다. 〈표 1〉은 전처리를 거친 최종 데이터의 변수 목록을 보여준다.

표 1. 수집된 데이터

종류	변수명	의미
종속변수	events	출루 / 아웃
독립변수	pitch_type	투구 유형
	release_speed	투구 속도
	release_pos_x	투구 가로 좌표
	release_pos_y	투구 높이 좌표
	release_pos_z	투구 세로 좌표
	release_spin_rate	투구 회전율
	release_extension	투수가 마운드에서 팔을 뻗은 거리
	spin_axis	투구 회전축
	zone	투구 구역
	stand	좌타 / 우타
	p_throws	좌완 / 우완
	strikes	스트라이크 개수
	balls	볼 수 개수
	pfx_x	투구 수평 움직임
	pfx_z	투구 수직 움직임
	plate_x	투구 가로 위치
	plate_z	투구 세로 위치
	outs_when_up	아웃 개수
	inning	이닝 수
	at_bat_number	현재 타석 번호
	pitch_number	현재 투수 번호
	bat_score	점수
	on_3b, on_2b, on_1b	주자 여부

3. 투수와 타자 유형별 군집변수 생성

MLB Stat Cast에서 수집한 상세 투구 데이터 이외에 투수와 타자 유형별 특성을 반영한 군집변수 생성을 위해 k-means로 군집을 만들어 특성을 파악했다. 선수마다 개인이 상대할 때 강한 모습을 보이거나 상대하기 어려워하는 유형이 존재한다. 이러한 특성을 반영하기 위한 과정 중 첫 번째로 k-means를 이용해 군집을 나누어 투수와 타자의 유형을 파악할 수 있는 변수를 생성했다. 이를 위해 본 연구에서는 투수와 타자 능력을 확인할 수 있는 투수와 타자의 성적 기록(타율, OPS, 방어율 등)을 k-means 입력 데이터로 이용했다(Marcou, 2020; 이장택, 2014). k-means 분석 전 군집의 개수를 정하기 위해 Elbow 기법을 사용했다. 그 결과 〈그림 2〉와 〈그림 3〉에서 확인할 수 있듯이 타자 군집의 수는 5개, 투수 군집의 수는 6개로 선정되었다. 그리고 선수별로 3년간의 기록의 평균을 사용하여 k-means 분석을 시행했다. k-means 모델 진행 과정을 살펴보면, 일단 min-max scale을 사용하여 데이터 스케일 조정 후, k-means를 실시했다. 군집분석 후, 군집 별 특성을 살펴보면, 타자 군집은 〈부록 표 3〉과 같이 나왔고, 투수 군집 같은 경우는 〈부록 표 1〉과 같은 결과를 보였다. 이렇게 투수와 타자 모두 합쳐 11개의 특성을 가진 유형으로 선수가 분류되었다.

타자 군집의 특성을 먼저 상세히 살펴보면, 1번째 타자 군집은 낮은 타율(0.220)을 기록했지만, 높은 볼넷율(9.1%), 높은 출루율(0.300), 준수한 장타

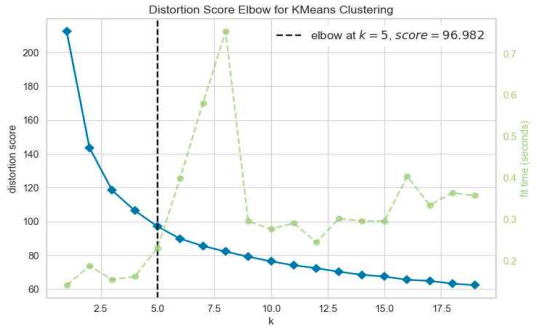

그림 2. Elbow method 결과: 타자 군집

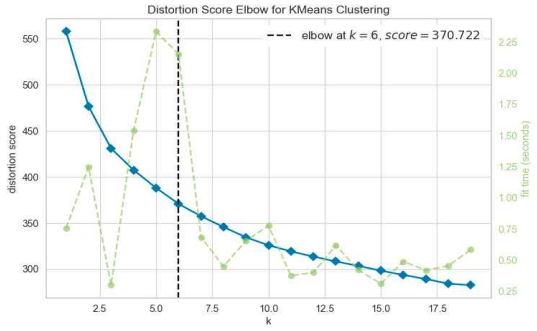

그림 3. Elbow method 결과 : 투수 군집

율(0.398)을 기록했다. 이로 보아 선구안이 좋은 출루형 타자 그룹으로 볼 수 있다. 이 군집에 해당하는 타자는 '최지만' 선수가 있다. 실제 최지만 선수는 다소 부족한 컨택트를 지녔지만, 좋은 선구안과 괜찮은 장타력을 가진 OPS 히터로 평가된다. 2번째 타자 군집은 꽤 괜찮은 볼넷율(7.5%)을 가졌지만, 메이저리그 평균보다 낮은 장타율(0.291), 평균보다 낮은 타율(0.195)를 기록했다. 이로 보아 선구안과 일발 장타력은 존재하지만, 잘 발휘 못한 타자 그룹으로 볼 수 있다. 이 군집에 해당하는 타자는 '쓰쓰고 요시토모'가 있다. 실제로 쓰쓰고 요시토모 선수는 선구안과 보통의 타격력은 있지만 93마일 이상의 빠른 공에 대한 대처가 전혀 되지 않아 메이저리그에서 지명 할당되었다. 3번째 타자 군집은 군집 중 가장 낮은 삼진율(20.3%)과 메이저리그 평균 이상의 타율(0.255), 높은 라인 드라이브 타구 비율(25%)을 기록하고 있다. 이로 보아 스윙 컨택률이 좋고 인플레이 타구를 많이 만들어 내고 질 좋은 타구를 많이 생산하는 그룹이다. 이 그룹에 속하는 타자는 '김하성'이 있다. 실제로 김하성 선수는 2023 시즌 메이저리그에서 좋은 타구 비율을 잘 생산하였고, 메이저리그 평균 타율을 넘어선 활약을 인정받아 골든 글러브를 수상했다. 4번째 타자 군집은 군집 중 제일 많은 타점(58)과 홈런 개수(59) 기록을 가지고 있다. 또한 높은 타율(0.258)을 기록하고 있다. 타점과 홈런을 많이 생산하는 파워형 타자 그룹으로 볼 수 있다. 이 타자 군집에는 '오타니 쇼헤이'가 있다. 실제로 오타니 선수는 메이저리그에서도 세 손가락 안에 드는 힘을 바탕으로 한 OPS 히터이고, 홈런 생산성 하나만큼은 현재 리그에서 애런 저지 다음 가는 수준으로 리그 최고의 홈런타자 중 한 명으로 평가받는다.

그 다음으로 투수 군집의 특성을 살펴보면, 첫 번째 투수 군집은 땅볼에 특화된 구종인 싱커 구사율(45.8%)이 높고, 땅볼 유도 비율은 51.4%로 전체 군집 중 가장 높다. 보통 MLB를 기준으로 투수 인플레이 타구 중 땅볼의 비율(GB%)이 50% 근처를 기록하면 땅볼 투수라고 말한다. 그래서 이 그룹은 땅볼 유도형 투수 그룹으로 볼 수 있다. 이 그룹에 속한 투수로는 펠릭스 페냐가 있다. 실제로 펠릭스 페냐 선수는 싱커와 슬러브성 브레이킹볼을 활용한 땅볼형 투수로 평가받고 있다. 두 번째 투수 군집은 높은 포심 구사율(51%), 높은 탈삼진율(24.6%), 군집 중 가장 빠른 직구 평균 구속(94.3마일), 많은 뛴 이닝 기록(60이닝)을 가지고 있다. 이 기록으로 보아 강점인 직구 구위로 승부 하는 삼진형 선발 투수 그룹으로 볼 수 있다. 이 그룹에 속한 투수로는 '맥스 슈어저'가 있다. 실제로 맥스 슈어저는 전형적인 탈삼진형 투수이고, 슈어저의 포심은 메이저리그 최고 수준의 회전수를 가졌고 2013년 최고의 패스트볼 1위로 선정되기도 했다. 3번째 투수 군집은 군집 중 평균 직구 구속(90.3 마일)과 평균 변화구 구속(78.8 마일)으로 느리지만, 변화구와 직구 구속 차이가 제일 크고, 볼넷 허용률(8.2%)이 가장 낮다. 이 기록을 보아 구속 차이와 제구력으로 승부하는 제구형 투수 그룹으로 볼 수 있다. 이 그룹에 속한 투수로는 대표적인 제구형 투수로 알려진 '류현진'이 있다. 4번째 투수 군집은 뛴 이닝 수가 31이닝으로 가장 적고, 방어율(7.51)은 가장 높다. 그리고 허용 타구 속도가 90마일로 가장 빠르다. 이 기록을 보아 등판할 때마다 정타 허용을 많이 하여 허용 타구 속도가 높고, 방어율이 높고, 메이저리그에 적응을 잘 못한 투수 그룹으로 보인다. 이 그룹에 속한 투수로는 '조던 야마모토' 선수가 있다. 실제 조던 야마모토 선수는 이후 부상과 부진에 시달리며 2023년 시즌 시작 전에 높은 방어율 등 좋지 않은 성적으로 인해, 현역 은퇴를 결정했다. 5번째 투수 군집은 적은 평균 이닝(39.5)

를 소화하고, 군집 중 가장 높은 삼진율(25.3%), 가장 낮은 방어율(3.95), 가장 낮은 빠른 타구 허용률(36.8%), 높은 직구 구사율(43%), 높은 슬라이더 구사율(46.5%)을 가지고 있다. 이를 보아 좋은 직구, 슬라이더 등을 주로 구사하여 타자들이 정타를 맞히기 힘든 투수 그룹으로 보인다. 실제로 이 군집에 속하는 투수는 '에드윈 디아즈' 선수가 있다. 쓰리쿼터에서 나오는 평균 99마일, 최고 103마일의 포심 패스트볼과 93마일의 슬라이더가 주무기인 전형적인 파이어볼러이고, 이에 따른 탈삼진율이 매우 높은 강속구 마무리 투수로 평가받고 있다. 6번째 투수 군집은 낮은 방어율(4.13), 많은 평균 이닝(63), 높은 삼진율 (24%), 모든 구종 구사 비율이 최소 10% 이상이다. 이 기록을 보아 다양한 구종을 던지는 에이스 선발 투수 그룹으로 보인다. 이 군집에 속하는 대표적인 투수는 '오타니 쇼헤이'다. 실제로 오타니 선수는 빠른 구속을 지닌 포심 패스트볼과 스위퍼, 스플리터 등의 유인구를 조화시킨 레퍼토리로 무수한 탈삼진을 뽑아내는 우완 강속구 투수이고, MLB 역사상 최초의 만장일치로 MVP 2번 수상을 기록한 에이스 투수로 평가받고 있다.

이처럼 군집화를 통해 투수와 타자의 특성을 살펴볼 수 있었다. 이 유형별 특성들을 반영하기 위해 각 군집에 해당하는 투수, 타자들의 군집 번호를 칼럼으로 생성했다. 해당 내용의 예시는 표 2에서 확인할 수 있다.

표 2. 군집 변수 생성

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

4. 데이터 병합과 스케일링, 오버샘플링

군집 변수 생성 후, 2020부터 2023년 메이저리그 투구 데이터와 k-means 실행 데이터의 군집변수 칼럼을 합하기 위해 player_id 기준으로 합병하여 데이터를 합치는 작업을 거쳤다. 데이터를 합한 후, 모든 특성의 범위를 맞춰주기 위해 표준화를 진행했다. 표준화 이후 종속 변수 내 아웃의 개수가 출루 개수보다 큰 비중으로 존재하여 종속 변수의 데이터 불균형 문제가 발생했다. 이를 해결하기 위해 SMOTE 기법을 사용했다. SMOTE는 대표적인 오버샘플링 방법의 하나로 분류 모형을 구축할 때 낮은 비율로 존재하는 그룹 데이터를 KNN 알고리즘 등을 활용하여 새롭게 생성하여 레이블 간 데이터 불균형을 해결할 수 있다(Chawla, 2002).

5. 선수 유형별 상대 출루율 변수 생성

선수의 유형을 통해 새로운 변수를 만들어내는 과정의 연장으로 투수와 타자 유형별 상대 간의 출루 정보를 입력으로 활용하기 위해, 클러스터링해서 얻은 군집변수와 종속 변수를 활용하여 타자 유형별 상대 출루율을 계산 후, 칼럼으로 만들어 활용했다. 예를 들면 P0에 해당하는 투수 군집과 B0에 해당하는 타자 군집이 맞붙었을 때 유형별 총 출루와 아웃 수를 카운트하여 전체 대결 별 출루의 비율을 계산한다. 이처럼 출루율을 계산 후, 'On_base_ratio_by_pitcher_vs_batter_type'이라는 이름으로 열을 생성하고, 해당 열에 계산된 출루율을 입력했다. 즉, 투수와 타자 유형별 상대 출루율의 의미를 지니는 파생 변수를 추가한 것이다. 이 변수 추가 이후, 실제 모델 성능 향상 있어 활용하기로 했다.

6. 예측을 위한 머신러닝 기법 선정

1) XGBoost

Chen과 Guestrin(2016)이 소개한 XGBoost는 선형 또는 트리 모델에서 발생하는 오버피팅을 해결하고 빅데이터에서 학습의 속도와 안정성을 개선하기 위해 고안된 알고리즘이다. 이 모델은 빅데이터에서 발생하는 복잡한 함수 및 결과 변수 간의 관계를 효율적으로 다루는 데 도움을 준다. 즉, 이것은 회귀, 분류 등이 가능한 부스팅 알고리즘 기반 모델인 XGBoost를 나타낸다. 부스팅 기법의 대표적 알고리즘인 Gradient Boost를 병렬 학습으로 구현한 알고리즘이기 때문에 성능과 효과가 매우 뛰어나다고 평가받는다.

2) LightGBM

LGBM 은 GBDT 기반 앙상블 모델 중 하나로, 내부 알고리즘을 사용하여 모델 구축에 사용되는 데이터 및 변수의 양을 줄여 계산 시간을 단축하고 성능을 향상시킨다. 이때 이용하는 알고리즘이 Gradient based one side sampling (GOSS)와 Exclusive feature bundling (EFB)이다. GOSS 는 모델 적용 시 입력 데이터를 정보 획득량의 절댓값 순으로 배열하는 알고리즘으로, 상대적으로 정보 획득이 큰 데이터와 작은 데이터에 서로 다른 가중치를 적용하여 모델 내 입력 변수를 선택적으로 활용하는 알고리즘이다. EFB는 고차원 변수를 하나의 변수로 그룹화하여 모델 구축에 사용되는 입력 변수를 줄이는 알고리즘이다(Ke et al, 2017). GOSS와 EFB를 적용해 메모리 사용량을 줄이면서 많은 양의 훈련 데이터를 학습할 때 빠르다는 장점이 있지만 오버피팅 이슈가 발생하기 쉬운 알고리즘이다. 오버피팅이 발생하기 쉬운 알고리즘이기 때문에 많은 양의 훈련 데이터가 필요하다.

3) CatBoost

앙상블 머신러닝모델 중 하나인 CatBoost모델은 gradient boosting machine (GBM)을 바탕으로 한 기계학습으로, ordered boosting을 사용하여 기존 GBM 으로 각 단계별 모형 구축 시, 그 시점의 예측 대상의 변수를 포함하여 모델의 오버피팅 가능성이 높아지는 문제를 해결하여 범주형 변수처리에 유용한 모델이다(Prokhorenkova, 2018).

7. 모형평가 기준

정확도(Accuracy)는 실제 데이터에서 예측 데이터가 얼마나 동일한지를 판단하는 지표이다. 또한, 직관적으로 모델이 얼마나 잘 예측한지 성능을 나타낼 수 있다 즉 정확도가 높을수록 예측 성능이 정확하다. 정확도의 계산식은 〈수식 1〉과 같다. 계산식에서는 TP는 True Positivie, TN은 True Negative, FP는 False Positive, FN은 True Negative이다.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

수식 1. 정확도

F_1 점수는 정밀도와 재현율의 가중조화 평균값이다. F1의 최댓값은 1, 최솟값은 0으로 F_1 점수가 높을수록 모형 예측 성능이 효과적이다. F_1 점수의 계산식은 〈수식 2〉와 같다. 계산식에서는 P는 정밀도 precision이고 R은 재현율 recall이다.

$$F_1 = \frac{2PR}{P + R}$$

수식 2. F_1 점수

III. 결 과

머신러닝 기법은 야구, 축구 등 스포츠 분야에서 성적 예측, 승부 예측 등에 활용하여 좋은 모델의 성능을 보여준 LGBM, CatBoost, XGBoost와 같은 다양한 머신러닝 기법을 비교하여 예측 모델을 구축에 사용했다(M. N. Razali, 2022; 한정섭, 2022). 이후 모델의 성능은 정확도, F_1 점수를 통해 평가했다.

모델의 성능을 높이기 위해 오버샘플링 전과 후, 다양한 변수 조합들을 고려했다. 오버샘플링 후, 투구에 관련된 모든 기록과 유형별 상대 출루율을 입력 변수로 넣었을 때 가장 모델 성능이 좋았다. 최종적으로 선택된 변수는 다음 〈표 3〉과 같다.

모델의 결과를 확인하기 위해 투구 데이터, 투수와 타자의 선수 유형별 대결 시의 출루율 등 총 전처리 된 변수들을 사용하였고, 학습 데이터와 검증 데이터를 8:2의 비율로 나눈 후, 출루 혹은 아웃을 예측하였다. 모델에는 LGBM, CatBoost, XGBoost를 활용하였다. 3가지 모델 모두 Random search CV를 활용하여 하이퍼 파라미터를 튜닝 후, 교차검증 타당성을 위해 5번의 교차 검증을 진행하여 모델 성능을 비교했다. 비교 결과, LGBM이 정확도 81.6%, F_1 점수가 0.81로 가장 우수한 결과를 보여줬다. 따라서 출루 혹은 아웃을 예측하는 최종 모델을 LGBM으로 선택하였다. 〈표 4〉는 최종 선택된 LGBM 모델의 하이퍼 파라미터를 보여주고, 〈그림 4〉는 모델 정확도 성능평가 결과를 나타낸 표이다.

표 4. 최종 변수

종류	변수명	의미
종속변수	events	출루 / 아웃
독립변수	pitch_type	투구 유형
	release_speed	투구 속도
	release_pos_x	투구 가로 좌표
	release_pos_y	투구 높이 좌표
	release_pos_z	투구 세로 좌표
	release_spin_rate	투구 회전율
	release_extension	투수가 마운드에서 팔을 뻗은 거리
	spin_axis	투구 회전축
	zone	투구 구역
	stand	좌타 / 우타
	p_throws	좌완 / 우완
	strikes	스트라이크 개수
	balls	볼 수 개수
	pfx_x	투구 수평 움직임
	pfx_z	투구 수직 움직임
	plate_x	투구 가로 위치
	plate_z	투구 세로 위치
	outs_when_up	아웃 개수
	inning	이닝 수
	at_bat_number	현재 타석 번호
	pitch_number	현재 투수 번호
	bat_score	점수
	on_3b, on_2b, on_1b	주자 여부
	On_base_ratio	선수별 출루율

표 3. 모델 하이퍼 파라미터

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

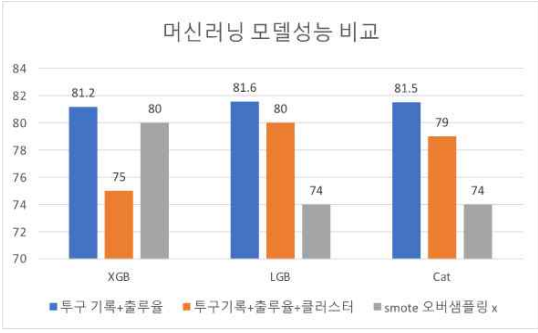

그림 4. 모델 성능

IV. 논 의

야구의 타격 예측에 관한 기존 연구는 간단한 통계적 기록을 활용하여 수행되었다. 그러나 이러한 연구들은 야구의 복잡한 상황을 충분히 반영하지 못했다. 그래서 기초 통계적 요소들 이외에 타격과 관련된 상세한 투구 기록을 입력으로 활용하였다. 또한, 손혁 (2004)의 연구에서 투수별 대응법을 강조한 것과 같이, 어떤 유형의 선수들 상대하느냐에 따라 출루의 결과가 달라질 것임을 확인했다. 이를 통해 투수, 타자 유형을 군집화 머신러닝을 활용하여 출루 아웃 예측 모델 개발을 진행했다.

본 연구는 스포츠 분야에서 선수의 성적이나 승부 예측 등에 활용하여 좋은 모델의 성능을 보여준 LGBM, CatBoost, XGBoost을 활용해 예측 모델을 구축했다. XGBoost는 부스팅 기법의 대표적 알고리즘인 Gradient Boost를 병렬 학습으로 구현한 알고리즘이기 때문에 성능과 효과가 매우 뛰어나다는 것으로 평가받고 있는 모델이다. LGBM은 GOSS와 EFB를 사용하여 메모리 이용을 줄이면서 많은 훈련 데이터를 학습할 때 빠른 이점이 있지만, 오버피팅이 발생하기 쉬운 알고리즘이다. CatBoost는 GBM을 바탕으로 한 기계학습 알고리즘으로, 기존 GBM 알고리즘이 Ordered Boosting을 사용할 경우, 각 단계에서 모델을 구축할 때 모델 오버피팅일 확률이 높아지는 이슈를 해결하여 범주형 변수를 다루는데 이점이 있는 모델이다.

인공지능 예측 모델 개발을 위해 2020년부터 2023년까지의 시즌에서 공의 무브번트, 릴리스 익스텐션, 던질 때 투수의 손 위치 좌표 등 상세한 투수 기록과 공격팀과 점수 차이, 주자 상황 등 투구 상황 기록 데이터를 수집하고 선수 유형별 군집 변수, 선수 유형별 상대 출루율로 파생 변수를 만든 후 활용하여 모델을 개발하였다. XGBoost는 정확도가 81.2%, F_1 점수가 0.805인 성능을 기록했고, LGBM은 정확도가 81.6%, F1 점수가 0.81인 성능을 기록했고, CatBoost는 정확도 81.5%, F1 점수가 0.81인 성능을 기록했다. 세 모델 중 정확도와 F1 점수가 가장 높은 것은 LGBM 모델이었다. 그래서 최종적으로 모델 성능이 가장 좋았던 LGBM으로 타자 출루 예측 모델로 선택되었다.

V. 결론 및 활용

본 연구에서는 MLB Stat Cast에서 수집된 투구 데이터를 기반으로, 기존의 단편적인 데이터 분석 방법을 넘어서 선수들의 성적과 플레이 스타일을 클러스터링하여 추가 변수로 활용함으로써 경기 투구 상황과 선수의 특성을 동시에 고려하여 정확도 81.6% 예측할 수 있는 LGBM 모델을 구축하였다.

이 연구의 예측 모델로 활용할 수 있는 것은 다음과 같다. 첫 번째로, 이 모델을 통해 팀 자원 파악, 선수기용 및 영입에 참고 할 수 있다. 팀 내 어떤 유형의 선수가 있는지 파악하여, 현재팀 전력으로 어떤 유형의 선수를 상대하는데 어려움이 있는

지 파악할 수 있다. 이를 통해 팀에 어떤 선수를 영입이 필요한지, 어떻게 훈련해야 하는지 전략을 세울 수 있다. 또한, 신인과 같은 적은 정보의 투수들을 상대해야 할 때, 모델을 활용하여 신인과 비슷한 유형의 투수를 찾아 그에 맞게 전략을 구상할 수도 있다.

두 번째는 MLB 리그 선수 유형별 경향을 파악하여 전략을 세우는데 활용할 수 있다는 점에 있다. MLB 2020부터 2022년까지의 출루 모델 결과, 땅볼 유도형 투수는 컨택트형 타자에 강하고, 파워형 타자 그룹에 약한 면모를 띠는 경향이 있다. 또한, 구위가 좋은 삼진형 투수는 컨택트형 타자에 강하고 파워형 타자 그룹에 약한 면모를 띠는 경향이 있다. 다양한 구종을 구사 가능한 투수는 홈런형 타자에게 강한 면모를 보였고, 컨택트형 타자에게는 약한 면모를 보였다. 이런 경향을 활용하여 교체 및 훈련 피드백 자료에 활용될 수 있을 것이다.

세 번째 활용은 선수 개인별 분석을 할 수 있다. 그 예시로 커쇼 선수의 2020에서 2022년 투구 데이터와 군집데이터를 활용하여 모델 적용 결과, 커쇼 선수의 출루 허용률을 살펴보니, 장타형 유형의 타자에게 강했으나, 컨택트형 타자에게는 약한 면모를 보였다. 이를 통해 커쇼 선수는 컨택트형 타자들에게 잘 대응할 수 있는 좀 더 세밀한 투구 전략을 세울 필요가 있다는 인사이트를 얻을 수 있었다. 이와 같이 개발된 모델을 이용해 선수 개인별 인사이트를 얻어 활용할 수 있다.

네 번째로, 예측 모델을 활용하여 사전 전략을 수립하는데 활용 할 수 있다. 구체적인 활용 계획을 살펴보면, 상대팀의 예상 라인업을 선별 후, 선발 투수의 투구 정보와 투수와 타자 유형을 입력하고, 타격결과를 예측한다. 이것으로 타자를 상대할 때 어느 위치에 어떤 공을 던질 때 출루 허용이 낮음을 모델로 확인하여 선수별 전략을 준비하는데 활용 할 수 있다. Otremba(2022)의 연구처럼 선수의 성향에 따라 어떤 상황에서 어떤 행동을 취할 것인지 시나리오를 다방면으로 파악해 놓는다면 투수가 까다로운 상황에 마주했을 때 코치진에서 직관적인 해결법을 제공할 수 있을 것이다.

본 연구가 가지는 의의는 새로운 관점의 야구 연구자료를 생성했다는 것에 있다. 새로운 방법론으로 투구 기록과 투수와 타자 군집을 활용하여 출루율을 예측하는 연구를 제시했다. 이는 스포츠 연구에 새로운 참고 자료로 활용될 수 있을 것이다.

참고문헌

1. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
2. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
3. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.
4. Marcou, C. (2020). Investigating Major League Baseball Pitchers and Quality of Contact through Cluster Analysis.
5. Razali, M. N., Mustapha, A., Mostafa, S. A., & Gunasekaran, S. S. (2022). Football Matches Outcomes Prediction Based on Gradient Boosting Algorithms and Football Rating

System. Human Factors in Software and Systems Engineering, 61, 57.
6. Nathan, A. M. (2012, October). What new technologies are teaching us about the game of baseball. In Proceedings of the Euromech Physics of Sports Conference.
7. Otremba Jr, S. E. (2022). SmartPitch: Applied Machine Learning for Professional Baseball Pitching Strategy. Unpublished Doctoral dissertation, Massachusetts Institute of Technology.
8. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.
9. 김혁주(2012). 한국 프로야구에서 출루 능력과 장타력이 득점 생산성에 미치는 영향. 한국데이터정보과학회지, 제23권 제6호, 1165-1174.
10. 문형우, 우용태, 신양우(2016). 한국 프로야구 경기에서 기대득점과 기대승리확률의 계산. 응용통계연구, 제29권 제2호, 321-330.
11. 박태신, 김재윤(2022). 머신러닝을 활용한 KBO 외국인 투수 재계약 예측 모형. 한국데이터정보과학회지, 제33권 제6호, 963-976.
12. 이장택(2014). 한국프로야구에서 타자능력의 측정. 한국데이터정보과학회지, 제25권 제2호, 349-356.
13. 이승훈, 최형준(2019). 미국 프로야구(MLB) 풀카운트 상황에서 투수의 구질, 구속 변화에 따른 투구 결과 분석. 한국체육과학회지, 제28권 제3호, 973-981.
14. 조선미, 김주학, 강지연, 김상균(2023). 머신러닝(XGBoost)기반 미국프로야구(MLB)의 투구별 안타 및 홈런 예측 모델 개발. 한국체육측정평가학회지, 제25권 제1호, 65-76.
15. 조형석(2021). MLB 타자들의 스윙존에 따른 스윙 선택 성향 분석. 미간행 석사학위논문, 명지대학교 기록정보과학전문대학원.
16. 최영환(2018). 4차 산업혁명형 ICT기술이 스포츠 분야에 미치는 기술 · 문화적 동향분석. 한국스포츠학회, 제16권 제3호, 1-12.
17. 황수웅(2023). 불확실성(uncertainty)을 고려한 스포츠 빅데이터 분석: Bayesian 추정과 Deep Learning을 활용한 프로야구 심판의 Ball/Strike 판정 평가 모델 개발. 미간행 박사학위논문, 서울대학교 대학원.
18. 한정섭, 정다현, 김성준(2022). 머신러닝을 활용한 빅데이터 분석을 통해 KBO 타자의 OPS 예측. 차세대융합기술학회논문지, 제6권 제1호, 12-18.

논문투고일 : 2023. 12. 31.
논문심사일 : 2024. 02. 02.
심사완료일 : 2024. 02. 05.