MLB 경기 및 날씨 데이터를 활용한 타구 위치 예측

신가연[1] · 김희련[2] · 이기찬[3] · 박태연[4] · 김선규[5] · 권민규[6,*]

[1]서강대학교 · [2]상명대학교 · [3]동양미래대학교 · [4]성결대학교 · [5]상지대학교 · [6]서울대학교

Batted Ball Location Prediction Using MLB Game Statistics and Weather Data

Ga-yeon Shin[1] · Hee-ryeon Kim[2] · Ki-chan Lee[3] · Tae-yeon Park[4] · Sun-kyu Kim[5] · Min-kyu Kwon[6,*]

[1]Sogang University · [2]Sangmyung University · [3]Dongyang-Mirae University · [4]Sungkyul University · [5]Sangji University· [6]Seoul National University

E-mail : gayeon0518@sogang.ac.kr / kcd05132@naver.com / dlrlcks0824@naver.com / viptaeyeon@gmail.com / edwin9903085@gmail.com / kmk0610@snu.ac.kr

요약

본 연구는 기존 안타 및 홈런, 출루 및 아웃을 예측하는 야구 데이터 분석에서 벗어나, 타구 위치(x, y 좌표)를 예측하는 모델을 제안한다. 이를 위해 2024년 MLB Statcast의 투구, 타자 및 타구 관련 데이터와 Meteostat API를 통해 수집한 타구 거리 및 방향에 영향을 미칠 수 있는 날씨 데이터(온도, 풍속, 강수량 등)를 기반으로 데이터셋을 구축하고, 예측 모델에 학습하였다. 그 결과 본 연구는 다양한 변수들을 기반으로 타구 위치를 예측하여, 데이터 기반의 효율적인 수비 시프트 최적화 및 투수 공 배합 전략 수립에 기여할 것으로 기대된다.

ABSTRACT

This study proposes a model to predict the batted ball location(x, y coordinates), moving beyond traditional baseball data analyses that focus on predicting hits and home runs, or on-base outcomes and outs. To achieve this, we collected pitch, batter, and batted ball data from the 2024 MLB Statcast, along with weather data (such as temperature, wind speed, and precipitation) from the Meteostat API, which are factors that may influence batted ball distance and direction. Using this dataset, we developed a predictive model. The results demonstrate that this study can predict batted ball location based on various variables, contributing to optimizing data-driven defensive shifts and establishing effective pitching strategies.

키워드

Machine learning, MLB, Prediction, Pitch/Batter data analysis, Batted ball Location

Ⅰ. 서 론

최근 프로스포츠 산업에서는 데이터 기반의 의사결정이 중요한 전략 요소로 자리 잡고 있다. 미국 메이저리그(MLB)에서는 2015년부터 Statcast 시스템을 도입하여, 타구 속도, 발사 각도 등 수많은 메트릭이 실시간으로 수집되고 있으며, 이는 경기 데이터를 활용한 다양한 분석 연구를 가능케 한다.[1] 이와 같은 환경은 스포츠 AI 분석 시장의 가능성을 높이며, 실시간 전략 의사결정 지원 도구로 발전하고 있다.

* 교신저자

기존의 야구 데이터 분석 연구는 주로 타자의 성과 또는 타구 결과 분류에 집중되어있다[2],[3]. 그러나 구체적인 타구의 낙하 좌표 중심의 예측 연구는 드물며, 타구 비행 및 낙하지점에 영향을 줄 수 있는 날씨 요인까지 고려하는 연구는 매우 부족하다. 이러한 배경 속에서 본 연구는 기존 연구와 차별화하여, 타구 및 경기 관련 데이터에 날씨 데이터를 함께 고려해 타구의 정밀한 위치(x, y 좌표)를 예측하는 모델을 제안하고자 한다.

Ⅱ. 데이터셋 구축

2.1 데이터 수집

2024년 MLB 정규 시즌 경기 데이터를 기반으로 총 세 가지 범주의 정보를 수집 및 통합하였다. 해당 데이터셋은 본 논문에서 HitTrackMLB라 명명하며, 총 109,971건의 타구 데이터를 대상으로 한다.

첫째, 타구 정보는 MLB 공식 데이터 플랫폼인 Baseball Savant의 Statcast 시스템 기반으로 수집되었으며, 타구 좌표(x, y), 투구 위치, 타자 타격 손 등의 타구 관련 정보를 포함한다.

둘째, 선수 정보는 각 타자 및 투수의 프로필 데이터와 스탯 지표를 중심으로 수집하였다.

셋째, 날씨 정보는 Meteostat API를 활용하여 각 경기장의 경기 시점의 기온, 풍속, 강수량 데이터를 수집하였다.

위의 3가지 관련 정보들을 타구 정보 기준으로 HitTrackMLB 데이터셋을 자체적으로 구축하였다.

표 1. HitTrackMLB 데이터셋 구성

역할	종류	변수명	의미
종속 변수	타구 정보	cx	타구 x좌표
		cy	타구 y좌표
독립 변수	선수 정보	player_age	타자 나이
		slg_percent	장타율
		isolated_power	순수 장타력
		babip	인플레이 타구 타율
		batter_hand	타자 타격 손
		pitcher_hand	투수 투구 손
	타구 정보	hit_dis	타구 비거리
		ball_type	타구 결과
		pitch_type	구종
		velocity	구속
		zone_num	투구 위치
		ball	볼
		strike	스트라이크
		exit_velocity	타구 속도
	기상 정보	temp	기온
		wind	풍속
		rain	강수량

2.2 전처리

타구 및 선수 관련 정형적인 데이터들의 경우 구종, 타구 결과 등 범주형 변수들은 라벨인코딩 방식을 활용하였고, 구속 및 타구 속도 등 연속형 변수들은 정규화를 사용하였다. 날씨 데이터 처리에서는 실외 구장의 경기들은 수집된 기상 데이터를 사용하였으며, 돔 구장의 경우에는 외부 기상관련 영향을 받지 않기에 강수량 및 풍속은 0으로 처리하고, 기온은 기본 22도로 처리하였다.

Ⅲ. 모델 학습 및 평가

3.1 모델 구축

타구 낙하 지점 x좌표(cx)와 y좌표(cy)를 각각 독립적인 목표 변수로 설정하고, MultiOutput-Regressor를 활용해 두 예측값을 통합하는 구조로 구성한 후 여러 회귀 모델에 적용하여 모델을 구축하였다.

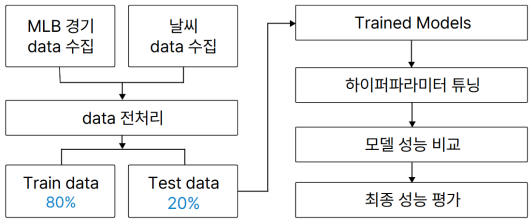


그림 1. 모델 개발 흐름 도식도

3.2 모델 비교 및 선정

모델 성능 평가지표로 평균 제곱근 오차(RMSE) 및 평균 절대 백분율 오차(MAPE)를 사용하여 각 모델 성능을 비교하였다.

표 2. 모델 성능 비교

모델	RMSE	MAPE(%)
Random Forest	46.3	16.2
XGBoost	42.4	13.4
LightGBM	44.2	14.2
DNN	49.8	18.3

모델 비교 결과, RMSE와 MAPE 모두 XGBoost 회귀 모델이 가장 우수한 예측 성능을 나타내어 최종 모델로 선정하였다.

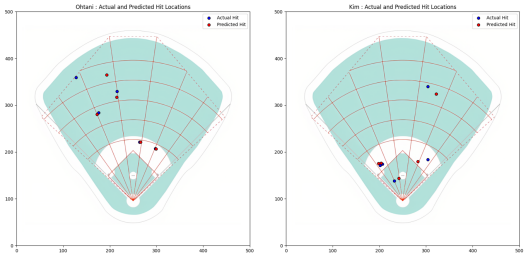
3.2 최종 모델 예측 타구 시각화

그림 2는 오타니 쇼헤이와 김하성 선수의 실제 타구 중 각 5개씩을 임의로 선정하여, 실제 낙하지점과 최종 예측 모델의 결과를 비교해 시각화한 것이다.

최종 모델의 예측 좌표와 실제 낙하지점 간의

평균 오차를 실제 거리로 환산한 결과 약 11.3m의 값을 보였다. 이는 MLB 선수의 평균 스프린트 속도(약 8m/s)[4]를 기준으로 수비수가 2초 이내에 커버할 수 있는 거리(15~17m)와 비슷한 수준의 예측 오차로, 모델이 실전 수비 전략 수립에 적용될 수 있는 실용적인 정밀도를 제공하는 것으로 해석할 수 있다.

그림 2. 예측 결과 시각화



Ⅳ. 결 론

본 연구는 기존 연구에서 주로 활용되던 타구, 선수 정보에 날씨 정보까지 통합적으로 반영하여 타구의 정확한 낙하 지점을 예측하는 모델을 제안하였다. 실제 경기 데이터를 기반으로 모델을 평가한 결과, 선수의 반응 속도와 커버 범위를 고려할 때 실전 적용 가능성이 충분한 수준의 예측 정확도를 확보한 것으로 판단된다. 따라서 본 연구 결과는 데이터를 기반으로 하는 효율적인 수비 시프트 최적화, 선수 타구 성향 분석 등에 활용할 수 있을 것으로 기대된다.

다만, 날씨 데이터의 한계로 인하여 타구별 개별적인 날씨 정보(특히 타구 시 풍량)에 대한 정보를 세부적으로 고려하지 못한 한계점은 남아 있어, 향후 연구에서 타구별 날씨 데이터를 반영하는 모델로 개선하고자 한다.

References

[1] Watkins, C., Berardi, V., & Rakovski, C., “Pitcher Effectiveness: A Step Forward for In-Game Analytics and Pitcher Evaluation”, Journal of Quantitative Analysis in Sports, 15, 4, 271-285, December 2019.

[2] 김명준, 정준호, “머신러닝(XGBoost)기반 미국프로야구(MLB)의 투구별 안타 및 홈런 예측 모델 개발”, 한국정보통신학회논문지, 26, 6, 1325-1333, 2022년 6월.

[3] 이승훈, 김준형, “머신러닝을 활용한 미국 프로야구의 투수 및 타자의 유형별 출루 및 아웃 예측 모델”, 한국융합학회 학술대회논문집, 제주, 292, 2022년 1월.

[4] MLB Advanced Media. Aaron Judge Statcast Hitting Data [Internet].Available : https://baseballsavant.mlb.com/leaderboard/sprint_speed?min_season=2024&max_season=2024&position=&team=&min=10