

머신러닝을 활용한 MLB 팀 성적 지표 분석 및 승률 예측 모델 연구

진병현¹, 권도윤², 이희화³

¹상명대학교 일반대학원 스포츠ICT융합학과 석사과정

²상명대학교 일반대학원 스포츠ICT융합학과 석사과정

³상명대학교 일반대학원 스포츠ICT융합학과 교수

iedl97@naver.com, hugokwonn@gmail.com, Leehh@smu.ac.kr

Performance Contribution Analysis of MLB Teams Using Machine Learning

Jeon Byeong-hyeon¹, Kwon Do-yoon², Lee hee-hwa³

¹Sangmyung University, Department of SportsICT Convergence

²Sangmyung University, Department of SportsICT Convergence

³Sangmyung University, Professor, Department of Sports Convergence

요 약

본 연구는 머신러닝 기반 MLB 데이터 분석을 통해 팀 성적에 기여하는 핵심 지표를 파악하고, 하위권 팀의 성과 향상을 위한 전략 시나리오를 제시하고자 한다. Statcast로부터 2020~2024 시즌의 투수 및 타자 데이터를 크롤링한 뒤, 상관관계 분석과 다중공선성 검증을 통해 주요 설명 변수를 선별하였다. 이후 랜덤 포레스트 모델과 주성분 분석(PCA)을 각각 수행하고, 하위 5개 팀을 대상으로 시나리오별 승률 변화를 예측하였다. 타자 지표는 OPS, R, BB%, K%, HardHit% 등을, 투수 지표는 K/9, BB/9, FIP, Pitching+, Location% 등을 활용하였다. 분석 결과, 팀별 약점 지표에 대한 소폭 개선만으로도 승률 상승 효과가 나타났다. 본 연구는 데이터 기반 야구 전략 수립에 실질적 기여를 할 수 있으며, 향후 KBO 적용 연구로도 확장 가능하다.

1. 서론

최근 메이저리그(MLB)를 비롯한 프로스포츠 산업은 고도화된 데이터 수집 기술과 분석 도구의 발전으로 인해 경기력 분석 및 전략 수립에 있어 정량적 접근의 중요성이 더욱 커지고 있다. 특히 타구 속도, 타구 각도, 회전수와 같은 스탯캐스트(Statcast) 기반의 세부 데이터가 널리 활용되면서, 팀 성과에 영향을 미치는 핵심 요소를 파악하고 이를 바탕으로 전략을 도출하려는 시도가 활발히 이루어지고 있다. 과거에는 선수 개인의 전통적인 기록에 의존한 분석이 주를 이루었으나, 최근에는 머신러닝 등 인공지능 기법을 통해 다변량 데이터를 통합적으로 해석하고, 변수 간 상호작용을 고려한 예측이 가능해졌다. 이러한 분석 방법은 팀 단위의 성과 예측이나 약점 진단에도 효과적으로 적용될 수 있으며, 특히 하위권 팀의 경우 제한된 자원 내에서의 전략적 개선 방향을 제시하는 데 매우 유용하다.

2. 연구방법 및 결과

2.1 데이터 개요

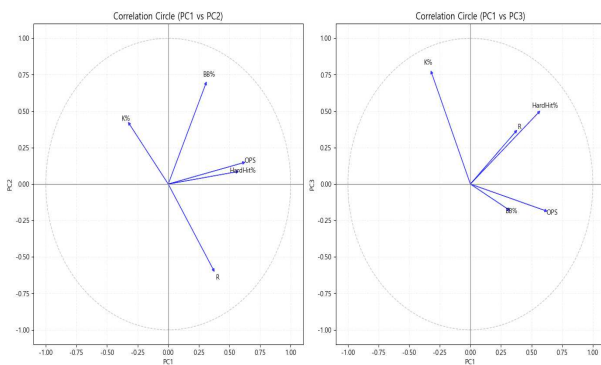
데이터는 MLB Statcast에서 2020년부터 2024년까지의 정규시즌 데이터를 크롤링하여 구축하였다. 타자 지표로는 OPS, R(득점), BB%(볼넷률), K%(삼진률), HardHit%(강한 타구 비율)를 포함하였고, 투수 지표로는 ERA(평균자책점), K/9(9이닝당 탈삼진), BB/9(9이닝당 볼넷), FIP(수비무관 평균자책점), Pitching+, Location%(투구 제구 정확도)를 수집하였다. 수집된 데이터는 팀 단위로 연도별 정리한 후, 변수 간의 중복성과 다중공선성을 제거하기 위해 상관관계 행렬 분석과 분산팽창지수(VIF) 검정을 수행하였다. 이를 통해 최종 분석에 사용할 변수를 선별하였다.

2.2 머신러닝 모델 구성 및 시나리오 분석

타겟 변수인 시즌별 팀 승률(Win_PCT)을 예측하기 위해 랜덤 포레스트 (Random Forest) 모델로 변수 중요도를 추출하였다.

주성분 분석(PCA)은 변수 간 구조적 관계를 해석하고 차원 축소를 통해 군집 특성을 시각화하기 위해 수행되었다. Correlation Circle 시각화로 변수 간 관계 해석을 실시하였다.

그림 1. PCA Correlation Circle.



시나리오 분석은 각 시즌별 승률 하위 5개 팀을 대상으로 수행하였다. 리그 평균과 비교하여 주요 지표 중 약점점으로 판단되는 항목에 대해 현실적인 범위 내에서 단계별 변화량을 설정하고, 해당 수치 변화에 따른 예측 승률을 산출하였다. 또한 모델의 예측 성능을 검증하기 위해 다양한 오차 지표를 활용하였다. MAE(Mean Absolute Error)는 실제 승률과 예측 승률 간의 절대적 평균 오차를 의미하며, MSE(Mean Squared Error)는 오차 제곱의 평균으로 큰 오차에 더 큰 패널티를 부여한다. RMSE(Root Mean Squared Error)는 MSE의 제곱근으로, 실제 승률과 예측 승률 간의 평균적인 차이를 직관적으로 해석할 수 있다. 마지막으로 MAPE(Mean Absolute Percentage Error)는 백분율 기준의 오차를 나타내어, 예측 정확도를 상대적으로 비교하는 데 유용하다. 본 연구에서 도출된 MAE는 약 0.044, MSE는 0.003, RMSE는 0.056, MAPE는 약 0.119로 제시된 시나리오 분석이 비교적 양호한 수준의 예측력을 확보했음을 확인할 수 있었다.

표 1. 시나리오 분석 결과

팀	시즌	변수	변화량	실제 승률	예상 승률	상승폭
BAL	2020	K/9	1	0.417	0.497	0.08
BAL	2020	ERA	-0.5	0.417	0.486	0.069
BOS	2020	BB/9	-0.5	0.4	0.42	0.02
DET	2020	K/9	0.6	0.397	0.415	0.018
DET	2020	ERA	-0.3	0.397	0.402	0.005
PIT	2020	ERA	-0.5	0.317	0.532	0.215
PIT	2020	BB/9	-0.3	0.317	0.407	0.09
TEX	2020	ERA	-0.5	0.367	0.488	0.121
TEX	2020	BB/9	-0.5	0.367	0.434	0.067
TEX	2020	K/9	1	0.367	0.403	0.036
ARI	2021	K/9	1	0.321	0.402	0.081
ARI	2021	BB/9	-0.5	0.321	0.389	0.068
BAL	2021	K/9	0.6	0.321	0.377	0.056
BAL	2021	BB/9	-0.5	0.321	0.373	0.052
TEX	2021	K/9	0.3	0.37	0.415	0.045
TEX	2021	ERA	-0.3	0.37	0.412	0.042
CIN	2022	ERA	-0.1	0.383	0.436	0.053
CIN	2022	BB/9	-0.5	0.383	0.434	0.051
COL	2023	K/9	1	0.364	0.388	0.024
COL	2023	BB/9	-0.5	0.364	0.386	0.022
NYM	2023	BB/9	-0.3	0.463	0.473	0.01
OAK	2023	BB/9	-0.5	0.309	0.376	0.067
MIA	2024	ERA	-0.1	0.383	0.413	0.03
OAK	2024	K/9	1	0.426	0.494	0.068
TOR	2024	K/9	1	0.457	0.517	0.06
TOR	2024	ERA	-0.5	0.457	0.501	0.044

그림 2. 실제 승률과 예측 승률 평가 지표

```

mae: 0.04465151515151516
mse: 0.00314554040404036
rmse: 0.05608511749154497
mape: 0.1193059349445082

```

3. 결론

본 연구는 MLB 하위권 팀을 대상으로 데이터 기반 시나리오 분석을 수행하여, 타자 및 투수 지표 개선 전략을 도출하였다. 머신러닝 기반 분석은 실무 전략 수립에 활용 가능하며, 향후 KBO 적용 및 선수 영입 전략 연구로 확장될 수 있다.

감사의글

본 연구는 문화체육관광부와 국민체육진흥공단의 지원을 받아 시행한 2025 스포츠산업 융복합대학원 사업의 결과물입니다.

참고문헌

- [1] 이기봉, 황수웅, "MLB 타자의 경기력 유형에 따른 팀 승률 변화 분석," 한국체육측정평가학회지, Vol. 25-4, 2023.
- [2] 안동수, 김지민, 손주찬, 김경외, "머신러닝을 활용한 미국 프로야구의 투수 및 타자의 유형별 출루 및 아웃 예측 모델," 한국체육과학회지, Vol. 33-1, 2024.
- [3] 이기봉, 황수웅, "서포트 벡터 회귀를 이용한 프로야구 팀의 승률 예측: 타자 경기력 유형 조합에 따른 설명 가능성을 중심으로," 한국체육측정평가학회지, Vol. 26-2, 2024.
- [4] 조선미, 김주학, 강지연, 김상균, "머신러닝(XGBoost)기반 미국프로야구(MLB)의 투구별 안타 및 홈런 예측 모델 개발," 한국체육측정평가학회지, Vol. 25-1, 2023.