KBO 팀시즌 득점 예측 모델 분석

1. 목표: 팀의 시즌 득점(R) 예측하기

KBO 팀의 시즌별 주요 타격 스탯(안타, 홈런, 볼넷) 등을 독립 변수로 사용하여, 해당 시즌의 총 득점 (R)을 예측하는 OLS(최소제곱법) 회귀 모델을 구축하고 그 성능 평가

2. 득점 예측 모델 구축

팀 득점 (R)을 종속 변수로 설정하고, 득점에 영향을 미칠 것으로 예상되는 주요 공격 스탯을 독립 변수로 사용하여 회귀 모델 설정

- 종속 변수: R(시즌 총 득점)

- 독립 변수: H(안타), HR(홈런), BB(볼넷), SB(도루), SO(삼진)

3. 모델 평가 및 정확도

OLS 회귀 분석 결과, 모델은 전반적으로 매우 높은 설명력을 보였음

모델 설명력(R-squared): 값이 0.938로 매우 높게 나타남. 이는 모델이 팀 득점 변동의 93.8%를 성공적으로 설명한다는 것을 의미

변수 유의성: H(안타), HR(홈런), BB(볼넷), SB(도루)는 P값(P>|t|)이 0.001이하로 나타나, 득점 예측에 매우 유의미한 영향을 미치는 변수임을 확인

반면 SO(삼진)의 P값은 0.733으로, 득점 예측에 통계적으로 유의미한 변수가 아닌 것으로 나타남

4. 최종 결과 분석: 예측 오차율(%)

모델의 예측값(Predicted_R)와 실제값(R)을 비교하여 '오차율(%)' 분석

전반적 분포: '오차율(%)분포' 히스토그램을 보면, 대부분의 오차율이 0%에 가깝게(왼쪽으로 치우침)에 분포하여 모델의 예측이 전반적으로 정확함을 알 수 있음.

가장 정확한 예측: 2024년 SSG (오차율 0.055%), 2021년 NC (오차율 0.063%) 등은 실제값과 예측값의 차이가 거의 없어 예측이 가장 정확했음.

가장 부정확한 예측: 반면, 2021년 키움(오차율 6.73%), 2021년 KIA(오차율 6.59%) 등은 모델의 예측과 실제 득점 간의 차이가 상대적으로 크게 나타남.

분석 참고사항: OLS 요약에서 조건수가 높게 나타나, 변수 간 '강한 다중공선성'이 존재할 수 있음이 지적됨.

5. 최종 결론 및 제언

단순히 팀 기록을 나열하는 것을 넘어, 어떤 스탯이 실제 득점 생산으로 이어지는지 통계적으로 증명함.

결론: 팀의 안타(H), 홈런(HR), 볼넷(BB), 도루(SB)는 시즌 총 득점을 예측하는 데 매우 유의미한 핵심 지표임이 확인됨. 반면, 삼진(SO)은 득점과의 직접적인 연관성이 유의미하지 않았음.

분석 제언:

모델 개선: 예측 오차가 컸던 특정 팀/시즌(예: 2021년 키움, KIA)을 아웃라이어로 처리하거나, 다중공선성 문제를 해결하여 모델을 더욱 정교화 할 수 있음.

활용 방안: 이 득점 예측 모델은 각 스탯이 득점에 얼마나 기여하는지 비교하여, 팀 공격력 구성에 있어 어떤 요소가 더 효율적인지 평가하는 기준으로 활용될 수 있음.