

야구선수 월별 성적 분석

1. 목표: 선수의 진짜 월별 기량 파악하기

프로젝트 목표: 구자욱 선수의 연도별/월별 타격 데이터를 분석하여, 각 월의 성적 변동 패턴을 파악하기

2. 평균의 함정

- 선수의 월별 타율을 분석해보니, 어떤 달은 4할이 넘는 맹타를 휘두르다가도 다른 달에는 1할 대에 그치는 등 기복이 심함
- 하지만 데이터를 자세히 보면 통계적 함정이 있는데, 부상 복귀나 시즌 초나 후반에 출전이 적었던 달에는 타수(AB)가 비정상적으로 적었음. 예로 들어서, 한 달에 10타수 4안타만 쳐도 타율은 4할로 기록되는데, 이는 한 달 동안 경기를 나온 100타수 30안타를 쳐서 3할인 타자보다 더 잘했다고 말하기 어려움

3. 문제 해결을 위한 접근: 만약 그랬더라면?

만약 선수가 부상 없이 정상적으로 타수를 소화했다면, 그 달의 성적은 어땠을까? 라고 생각해서 데이터를 보정함

- 타수 보정: 먼저 선수의 월별 타수를 모두 계산 한 후, modified_AB_if_smaller라는 변수로 새로운 기준을 만들었음. 실제 타수가 월별 평균보다 적으면, '평균 타수'를 부여하고, 그렇지 않으면 실제 타수를 그대로 사용
- 타율 시뮬레이션: 타수가 보정된 달에 한해, 해당 월 커리어 평균 타율과 표준편차를 기반으로 가상 타율을 랜덤하게 생성, 이 가상 타율과 보정된 타수를 곱해 '가상 안타 수'를 계산하고, 이를 다시 보정된 타수로 나누어 최종 시뮬레이션 조정 타율 확정

4. 최종 결과 분석

- 데이터 유출이나 잘못된 전처리 순서 같은 문제는 없었지만, 표본 크기가 작은 데이터가 어떻게 분석 결과를 왜곡할 수 있는지 명확히 확인
- 타수가 적었던 달의 비정상적으로 높거나 낮은 타율이, 시뮬레이션을 통해 선수의 원래 기량에 가까운 수치로 보정

5. 최종결론 및 제언

- 단순히 기록된 타율만 보는 것은 선수의 기량을 오해하게 만들 수 있다. 따라서 특히 부상이나 기타 이유로 출전이 적었던 달의 성적은 타수를 보정하고 시뮬레이션을 통해 재평가할 때 더 합리적인 해석이 가능
- 분석 제언

선수 평가의 새로운 기준: 보정된 '시뮬레이션 조정 타율'은 연봉 협상이나 다음 시즌 성적 예측 모델을 만들 때, 기존 타율보다 더 안정적이고 신뢰도 높은 변수로 활용

확장가능성: 이 분석 방법론을 타율(AVG) 뿐만 아니라 출루율(OBP), 장타율(SLG) 등 다른 비율 기반 스탯에도 적용하여 선수의 종합적인 능력을 평가

미래예측: 이 시뮬레이션 데이터를 기반으로 미래 시즌의 월별 성적 등락을 예측하는 시계열모델 개발가능