

변수 선택 관련 이슈

1. 승리와 거의 사실상 같은 말인 변수 제거

- Rd(0.9387의 상관관계, 점수차이를 의미)
- Pitch_wins(1.0000의 상관관계, 투수 승리를 의미, 사실상 승리와 똑같은 말)
- L(패배, 승리와 거의 유사한 말)

2. 다중공산성(변수 간 상관관계 생기는 변수들, 정보가 겹치는 변수를 의미하는듯) 제거

- Ops, r, avg_era 제거

3. 결론

1,2에 나온 사실 반영해서 6개 변수를 추려서 일단 모델에 돌려봐야 할듯

변수 (Variable)	VIF Value	카테고리	설명
ra	1.15	수비/투구	상대에게 허용한 실점
slg	4.90	타격 효율	장타율
total_hr	3.01	타격 결과	팀 홈런 수
total_bb	2.21	타격/출루	팀 볼넷 수
pitch_so	2.47	투구 효율	투수 탈삼진 수
pitch_bb	2.30	투구 효율	투수 볼넷 (피 볼넷) 수
변수 (Variable)	VIF Value	카테고리	설명
ra	1.15	수비/투구	상대에게 허용한 실점
slg	4.90	타격 효율	장타율
total_hr	3.01	타격 결과	팀 홈런 수
total_bb	2.21	타격/출루	팀 볼넷 수
pitch_so	2.47	투구 효율	투수 탈삼진 수
pitch_bb	2.30	투구 효율	투수 볼넷 (피 볼넷) 수