박사학위논문

한국프로야구 골든글러브 수상자 예측모델 비교

Comparison of prediction models for Korean professional baseball Golden Glove winners

제 출 자 : 권 순 규

지도교수 : 최 형 준

2023

체육학과

체육학 전공

단국대학교 대학원

한국프로야구 골든글러브 수상자 예측모델 비교

Comparison of prediction models for Korean professional baseball Golden Glove winners

이 논문을 박사학위논문으로 제출함

2023년 12월

단국대학교 대학원
체육학과
체육학 전공

권순규

권 순 규의 박사학위 논문을
합격으로 판정함

심 사 일 : 2023. 12. 01.

심사 위원장 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

단국대학교 대학원

(국문초록)

한국프로야구 골든글러브 수상자 예측모델 비교

단국대학교 대학원 체육학과

체육학 전공

권 순 규

지도교수 : 최 형 준

야구는 타 종목에 비해 기록이 다양하고 완벽하게 보존되며, 경기에서 일어난 모든 기록들을 통해 경기내용 및 경기결과의 복기가 가능하다. 이로 인해 야구는 흔히 기록의 스포츠라 불린다. 이처럼 다양한 기록이 존재하는 야구에서 각 포지션별로 최고의 선수에게 주어지는 “골든글러브”상이 존재한다. 현재 골든글러브 수상은 취재기자, 중계 PD, 해설위원 등 미디어 관계자를 대상으로 투표를 진행하다보니 포지션별 명확한 기준 및 중요 변인을 알아보는데 한계가 있다. 따라서 본 연구에서는 2003년~2022년까지 한국프로야구 골든글러브 후보 및 수상자의 기록을 기반으로 로지스틱 회귀분석과 머신러닝을 활용한 골든글러브 수상자 예측모델을 설계하고 설계된 예측모델의 성능을 비교・분석하여 골든글러브 수상자 예측에 적합한 모델을 알아보는데 목적이 있다. 또한, 각 포지션별로 수상에 중요한 영향을 미치는 변인을 도출하고자 한다. 이 연구의 목적을 달성하기 위해 로지스틱 회귀분석

과 서포트 백터 머신(Support vector machine), 랜덤포레스트(Random Forest), XGboost 모델을 설계하고 각 모델별 하이퍼 파라미터를 제시하였다. 값을 표준 점수화 하여 나타내는 zscoring값과 최대-최소 정규화를 하여 나타내는 minmax값을 사용하여 각각의 모델을 두가지 형태로 나타냈으며, 모델별로 최적화 변인 탐색 후 성능평가를 실시하였다.

첫째, 한국프로야구 골든글러브 예측모델을 설계하는데 있어 로지스틱 회귀분석 모델에서는 L1, L2, elasticnet이 커널(kernel)로 사용되었고, 서포트 백터 머신 모델에서는 rbf, poly가 커널(kernel)로 사용되었으며, 비선형 모델로서 중요 변인은 탐색하지 못하였다. 랜덤포레스트 모델에서는 gini, entropy가 준거(criterion)로 사용되었으며, XGBoost 모델에서는 exact, approx, hist가 준거(criterion)로 사용되었다. F1 score가 높아지도록 하기 위하여 변인을 하나씩 제거하는 방식으로 진행하였고, 각 모델에서 포지션별 사용 변인은 모두 다르게 선정되었다.

둘째, 머신러닝 예측모델의 예측 성능을 비교한 결과 각 포지션별 차이는 존재하지만 서포트 백터 머신 모델 2와 XGBoost 모델의 예측 정확도, 그리고 F1 score가 높게 나타났으며, 로지스틱 회귀분석 모델과 랜덤포레스트 모델의 정확도와 F1 score는 상대적으로 낮게 나타났다. 전체적으로 zsocring으로 표준화 한 모델 1보다 minmax로 표준화 한 모델 2의 예측 능력이 뛰어나게 나타났다.

결론적으로 골든글러브 후보 및 수상자의 기록을 기반으로 골든글러브 수상자 예측이 가능하였으며, 각 모델별 중요 변인을 탐색할 수 있었다. 또한, 성능평가를 통해 XGBoost 모델 2가 가장 적합한 모델로 나타났다.

따라서, 골든글러브 수상자를 예측하기 위해서는 minmax를 사용한 XGBoost 모델 2를 활용하여 예측하는 것이 바람직하며, 추후 수비기록을 포함하여 예측한다면 보다 뛰어난 결과가 나타날 것이라 사료된다.

주제어 : 한국프로야구, 세이버메트릭스, 골든글러브 수상자 예측, 로지스틱 회귀분석, 머신러닝, 예측기법

목 차

국문초록 ····· i
목 차 ····· iii
표 목 차 ····· vi
그림목차 ····· ix

Ⅰ. 서론 ····· 1
1. 연구의 필요성 ····· 1
2. 연구 목적 ····· 5
3. 연구 문제 ····· 5
4. 용어의 정의 ····· 6

Ⅱ. 이론적 배경 ····· 10
1. 한국프로야구 골든글러브 ····· 10
1) 골든글러브 선정 ····· 10
2) 연도별 골든글러브 선정 조건의 변화 ····· 10
3) 미국프로야구 골든글러브와의 차이점 ····· 11
2. 정량적 자료분석의 중요성 ····· 12
1) 세이버메트릭스의 정의 ····· 12
2) 스포츠 경기분석의 통계적 접근 ····· 13
3. 머신러닝 기법 ····· 15
1) 머신러닝의 정의와 기원 ····· 15
2) 머신러닝의 종류 ····· 15
3) 학습의 원리 ····· 16
4) 머신러닝과 수학 ····· 17
4. 로지스틱 회귀분석 ····· 18
5. 서포트 백터 머신(Support vector machine) ····· 20
6. 랜덤포레스트 ····· 21
7. XGBoost ····· 23

Ⅲ. 연구 방법 ·····25
1. 연구대상 ·····25
2. 자료수집 도구 ·····28
3. 연구절차 ·····30
4. 자료처리 ·····32

Ⅳ. 연구결과 ·····36
1. 골든글러브 후보 및 수상자 간 포지션별 기록 비교 ·····36
1) 투수 변인별 기록 ·····37
2) 1루수 변인별 기록 ·····39
3) 2루수 변인별 기록 ·····41
4) 3루수 변인별 기록 ·····43
5) 유격수 변인별 기록 ·····45
6) 포수 변인별 기록 ·····47
7) 외야수 변인별 기록 ·····49
8) 지명타자 변인별 기록 ·····51
2. 로지스틱 회귀분석 및 머신러닝 예측모델 분석 결과 ·····53
1)로지스틱 회귀분석 및 머신러닝 예측 모델별 최적 변수 결정 ·····53
(1) 로지스틱 회귀분석 ·····53
(2) 서포트 백터 머신 ·····66
(3) 랜덤포레스트 ·····72
(4) XG부스트(XGBoost) ·····88
2) 로지스틱 회귀분석 및 머신러닝 예측 모델별 성능평가 결과 ····100
(1) 로지스틱 회귀분석 모델 포지션별 성능평가 결과 ·····100
(2) 서포트 백터 머신 모델 포지션별 성능평가 결과 ·····101
(3) 랜덤포레스트 모델 포지션별 성능평가 결과 ·····102
(4) XGBoost 모델 포지션별 성능평가 결과 ·····103

Ⅴ. 논의 ·····104

Ⅵ. 결론 및 제언 ·····110
1. 결론 ·····110
2. 제언 ·····112

참고문헌 ·····113

Abstract ·····118

표 목 차

표 1. 2003년～2022년 골든글러브 후보 수 ····· 25
표 2. 프로야구 타자기록(기본 기록, 세이버 메트릭스 기록) ····· 26
표 3. 프로야구 투수기록(기본 기록, 세이버 메트릭스 기록) ····· 27
표 4. 골든글러브 수상자 예측 모델 개발 연구 절차 ····· 31
표 5. 투수 세이버 메트릭스 기록 변인 설명 ····· 32
표 6. 투수 기본기록 변인 설명 ····· 33
표 7. 타자 세이버 메트릭스 기록 변인 설명 ····· 34
표 8. 타자 기본기록 변인 설명 ····· 34
표 9. 로지스틱 회귀분석 및 머신러닝 모델 ····· 35
표 10. 투수 기본 기록 기술통계 ····· 37
표 11. 투수 세이버 메트릭스 기록 기술통계 ····· 38
표 12. 1루수 기본 기록 기술통계 ····· 39
표 13. 1루수 세이버 메트릭스 기록 기술통계 ····· 40
표 14. 2루수 기본 기록 기술통계 ····· 41
표 15. 2루수 세이버 메트릭스 기록 기술통계 ····· 42
표 16. 3루수 기본 기록 기술통계 ····· 43
표 17. 3루수 세이버 메트릭스 기록 기술통계 ····· 44
표 18. 유격수 기본 기록 기술통계 ····· 45
표 19. 유격수 세이버 메트릭스 기록 기술통계 ····· 46
표 20. 포수 기본 기록 기술통계 ····· 47
표 21. 포수 세이버 메트릭스 기록 기술통계 ····· 48
표 22. 외야수 기본 기록 기술통계 ····· 49
표 23. 외야수 세이버 메트릭스 기록 기술통계 ····· 50

표 24. 지명타자 기본 기록 기술통계 ····· 51
표 25. 지명타자 세이버 메트릭스 기록 기술통계 ····· 52
표 26. 로지스틱 회귀분석 파라미터 설명 ····· 53
표 27. 로지스틱 회귀분석 모델 1 투수 및 포수 Hyper parameter ····· 54
표 28. 로지스틱 회귀분석 모델 1 내야수 Hyper parameter ····· 54
표 29. 로지스틱 회귀분석 모델 1 외야수 및 지명타자 Hyper parameter · 55
표 30. 로지스틱 회귀분석 모델 2 투수 및 포수 Hyper parameter ····· 59
표 31. 로지스틱 회귀분석 모델 2 내야수 Hyper parameter ····· 60
표 32. 로지스틱 회귀분석 모델 2 외야수 및 지명타자 Hyper parameter · 60
표 33. 서포트 백터 머신 파라미터 설명 ····· 66
표 34. 서포트 백터 머신 모델 1 투수 및 포수 Hyper parameter ····· 66
표 35. 서포트 백터 머신 모델 1 내야수 Hyper parameter ····· 67
표 36. 서포트 백터 머신 모델 1 외야수 및 지명타자 Hyper parameter ··· 68
표 37. 서포트 백터 머신 모델 2 투수 및 포수 Hyper parameter ····· 69
표 38. 서포트 백터 머신 모델 2 내야수 Hyper parameter ····· 70
표 39. 서포트 백터 머신 모델 2 외야수 및 지명타자 Hyper parameter ··· 71
표 40. 랜덤포레스트 파라미터 설명 ····· 72
표 41. 랜덤포레스트 모델 1 투수 및 포수 Hyper parameter ····· 72
표 42. 랜덤포레스트 모델 1 내야수 Hyper parameter ····· 74
표 43. 랜덤포레스트 모델 1 외야수 및 지명타자 Hyper parameter ····· 75
표 44. 랜덤포레스트 모델 2 투수 및 포수 Hyper parameter ····· 80
표 45. 랜덤포레스트 모델 2 내야수 Hyper parameter ····· 81
표 46. 랜덤포레스트 모델 2 외야수 및 지명타자 Hyper parameter ····· 83
표 47. XGBoost 파라미터 설명 ····· 88

표 48. XGBoost 모델 1 투수 및 포수 Hyper parameter ····· 88
표 49. XGBoost 모델 1 내야수 Hyper parameter ····· 89
표 50. XGBoost 모델 1 외야수 및 지명타자 Hyper parameter ····· 90
표 51. XGBoost 모델 2 투수 및 포수 Hyper parameter ····· 94
표 52. XGBoost 모델 2 내야수 Hyper parameter ····· 94
표 53. XGBoost 모델 2 외야수 및 지명타자 Hyper parameter ····· 95
표 54. 각 포지션별 로지스틱 회귀분석 모델별 성능평가 결과표 ····· 100
표 55. 각 포지션별 서포트 백터 머신 모델별 성능평가 결과표 ····· 101
표 56. 각 포지션별 랜덤포레스트 모델별 성능평가 결과표 ····· 102
표 57. 각 포지션별 XGBoost 모델별 성능평가 결과표 ····· 103

그림목차

그림 1. 로지스틱 회귀분석 모델 ····· 18
그림 2. Support vector machine 모델 ····· 20
그림 3. 랜덤포레스트 모델 ····· 22
그림 4. XGBoost 모델 ····· 23
그림 5. 골든글러브 후보 검색 엔진 ····· 28
그림 6. 골든글러브 후보 및 수상선수 기록 정리 예시 ····· 29
그림 7. 로지스틱 회귀분석 모델 1 투수, 포수 변인 중요도 ····· 56
그림 8. 로지스틱 회귀분석 모델 1 1루수, 2루수, 3루수 변인 중요도 ····· 57
그림 9. 로지스틱 회귀분석 모델 1 유격수, 외야수, 지명타자 변인 중요도 ····· 58
그림 10. 로지스틱 회귀분석 모델 2 투수, 포수 변인 중요도 ····· 62
그림 11. 로지스틱 회귀분석 모델 2 1루수, 2루수, 3루수 변인 중요도 ····· 63
그림 12. 로지스틱 회귀분석 모델 2 유격수, 외야수, 지명타자 변인 중요도 ····· 64
그림 13. 랜덤포레스트 모델 1 투수, 포수 변인 중요도 ····· 77
그림 14. 랜덤포레스트 모델 1 1루수, 2루수, 3루수 변인 중요도 ····· 78
그림 15. 랜덤포레스트 모델 1 유격수, 외야수, 지명타자 변인 중요도 ····· 79
그림 16. 랜덤포레스트 모델 2 투수, 포수 변인 중요도 ····· 84
그림 17. 랜덤포레스트 모델 2 1루수, 2루수, 3루수 변인 중요도 ····· 85
그림 18. 랜덤포레스트 모델 2 유격수, 외야수, 지명타자 변인 중요도 ····· 86
그림 19. XGBoost 모델 1 투수, 포수 변인 중요도 ····· 91
그림 20. XGBoost 모델 1 1루수, 2루수, 3루수 변인 중요도 ····· 92
그림 21. XGBoost 모델 1 유격수, 외야수, 지명타자 변인 중요도 ····· 93
그림 22. XGBoost 모델 2 투수, 포수 변인 중요도 ····· 96

그림 23. XGBoost 모델 2 1루수, 2루수, 3루수 변인 중요도 ····· 97
그림 24. XGBoost 모델 2 유격수, 외야수, 지명타자 변인 중요도 ····· 98

Ⅰ. 서 론

1. 연구의 필요성

야구는 타 종목에 비해 다양한 기록이 존재하고(황서영, 2006), 완벽하게 보존되며 경기에서 일어난 모든 기록들을 통하여 경기내용 및 경기결과의 복기가 가능하다(최경호, 2009). 이로 인해 야구는 흔히 기록의 스포츠라 불린다(김차용, 2001). 야구기록은 크게 두 가지로 구분할 수 있는데, 기본 기록과 통계적으로 만들어진 세이버 메트릭스로 나눌 수 있다. 기본 기록으로는 타자들의 타격능력을 나타내는 타격기록, 투수들의 투구능력을 나타내는 투수기록, 수비수들의 수비능력을 나타내는 수비기록이 존재한다(권순규, 이규원, 최형준, 2019). 기본 기록을 보게 되면 안타 수, 타점 수, 득점 수, 홈런 수, 탈 삼진수, 이닝 수, 실책 수, 보살 수 등의 횟수를 나타내는 기록과 타율 출루율, 장타율, 방어율, 피안타율, 수비율, 실책율 등의 확률을 나타내는 기록이 있다.

세이버 메트릭스는 미국야구연구학회(SABR: Society for American Baseball Research)의 회원인 빌 제임스(Bill James)가 미국야구연구학회의 SABR과 계량적 분석을 나타내는 metrics의 합성어인 세이버 메트릭스를 처음 사용하였으며(SABR, 2016), 세이버 메트릭스는 야구의 기록을 통계학적/수학적으로 분석하여 많은 지표가 만들어졌다(안현호, 2015). 초기에는 야구계의 별다른 관심을 받지 못했지만 2002년 미국프로야구(Major League Baseball) 오클랜드 애슬레틱스(Oakland Athletics)의 빌리 빈(Billy Beane)단장이 출루율 + 장타율 즉 OPS가 승리가 밀접한 영향을 미친다는 분석을 바탕으로 OPS형 타자를 영입하여 팀을 재구성하였는데 시즌 막판 20연승과 더불어 4년 연속 포스트시즌에 진출시킴으로써(양도업, 조은형, 배상우, 정상원, 2015) 현재 야구 기록의 꽃이라 불리는 세이버 메트릭스 이론이 주목 받기 시작하였다(승희배, 강기훈, 2012). 대표적인 세이버 메트릭스 지표로는 OPS(on-base Percentage Plus Slugging percentage, 출루율+장타율), BABIP(Batting average on Ball In Play, 인플레이 타구에 대한 타율), RC(Run Created, 득점창출 능력), IsoP(Isolated Power, 순수장타율)등의 타격지표와 WHIP(Walk and HIt pre Innings Pitched, 이닝당 출루 허용율), FIP(Fielding Independent

Pitching, 수비 무관 평균자책점)등의 투수지표와 타격, 수비, 투구기록 모두 합쳐 통계적으로 나타낸 수치인 WAR(Wins Above Replacement, 대체선수 대비 승리 기여도), WPA(Win Probability Added, 승리 확률 기여도) 등이 있다. 한국프로야구는 “야구인에게 자립과 복지를, 어린이에게 꿈을, 젊은이에게 젊음과 낭만을 그리고 국민에게는 건전한 여가선용과 즐거운 주말을 선사한다.” 는 슬로건으로 1982년 정치적이 목적을 배경으로 출범하였다(강준만, 2009; 전용배, 김애랑, 2011). 각 지역을 연고로 하는 6개 구단(삼성 라이온즈, 롯데 자이언츠, 삼미 슈퍼스타즈, 해태 타이거즈, MBC 청룡, OB베어스)로 시작되어 1987년 빙그레 이글스 창단, 1991년 쌍방울 레이더스가 합류하였고(양도업, 2016) 2013년 NC 다이노스, 2015년 KT 위즈의 창단으로 현재 10구단 체제로 운영 중이다. 2006년 WBC(World Baseball Classic) 4강, 2008년 베이징 올림픽 금메달, 2009년 WBC 준우승, 2015 WBSC 프리미어 12 우승함으로써 새로운 야구팬들을 대거 양산시켰고(김욱기, 2011) 그 결과로 많은 팬들의 사랑을 받으며 국내 최고의 인기 스포츠로 등극하였다. 한국프로야구위원회는 미국 메이저리그, 일본 프로야구와 더불어 세계 3대 프로야구 리그의 명성에 맞게 최근 출범 초기부터 이루어진 모든 기록 정리를 하였다(한국야구위원회. 2020). 2001년 이전 기록의 데이터화를 실시한 결과 전준호 선수의 도루기록 감소, 이강철 선수의 탈삼진 수 증가, 한용덕 선수의 탈삼진 수 증가 등 한국프로야구 초창기의 모든 기록을 완벽하게 정리 하였다(배중현 2020). 이처럼 한국프로야구는 야구 기록의 중요성 인지하고 초창기 미흡했던 기록을 완벽하게 구현하게 되었다.

야구 기록은 현재 한국프로야구의 선수 평가, 연봉 산정 등에 막대한 영향을 미치고 있음을 알 수 있다. 또한 프로야구 경기를 TV중계로 보면 기본 기록과 많은 세이버 메트릭스 지표들이 노출되고 있다. 예를 들면 투수가 경기에 등판하게 되면 승리 수, 방어율, 피안타 수 등의 기본 기록과 WHIP(이닝당 출루 허용율), K/BB(삼진 하나당 볼넷 비율) 등의 기록을, 타자가 경기에 출전하게 되면 안타 수, 타율, 타점, 득점 등의 기본 기록과 OPS(출루율 + 장타율), BABIP(인플레이 타구에 대한 타율) 등의 기록이 나오며, 각 팀별 또는 포지션별 선수의 순위를 나타낸 WAR(대체선수 대비 승리기여도)이 많이 노출되고 있다. 현대야구에서 기록에 대한 중요도가 높아지면서 한국프로야구위원회는 공식 기록을 토대로 다양한 시상을 실시하고 있다. 타자의 경우 타율상, 최다 안타상, 홈런상, 도루상, 득점상, 타점상, 출루율상, 장타율상을 시상하고 있으며, 투수의 경우 승리상, 탈삼진상, 평균자책점상, 홀드상, 세이브상, 승

률상을 시상하고 있다(한국야구위원회, 2022). 이를 토대로 최우수 선수상(MVP), 신인상, 그리고 각 포지션별 최고의 선수를 뽑는 “골든글러브”를 시상하고 있다.

골든글러브는 한국프로야구 144경기의 정규리그에서 가장 뛰어난 활약을 펼친 선수에게 프로야구 취재 기자단에서 투표하여 뽑는 상이다. 야구 종주국인 미국프로야구(MLB)의 경우 각 포지션별 최고의 수비선수를 뽑는 “골든글러브”상과 각 포지션별 최고의 공격선수를 뽑는 “실버슬러거(Silver Slugger Award)”상으로 구분되어 있다. 30개 구단 감독과 각 팀당 최대 6명의 코치진의 투표와 미국 야구 연구 협회의 수비 지표를 종합하여 골든글러브 수상자를 결정한다. 한국프로야구에서 타격 능력을 고려한 ‘실버슬러거’와 같은 수상이 없다보니 ‘골든글러브’의 수상자가 수비능력이 주가 되기보다 각 포지션의 최고의 공격 지표를 기록한 선수가 골든글러브를 받고 있다. 투수, 포수, 1루수, 2루수, 3루수, 유격수, 외야수 3명, 지명타자까지 10명의 선수가 골든글러브를 수상하고 있다. 투수의 경우 타이틀 홀더, 규정이닝 이상, 10승 이상, 30세이브 이상, 30홀드 이상 중 하나라도 해당하면 후보가 될 수 있으며, 지명타자는 타이틀홀더 또는 297타석 이상을 출전하게 되면 후보가 될 수 있다. 야수 및 포수의 경우는 타이틀홀더 또는 해당 포지션에서 수비를 720이닝 이상 출전하면 후보가 될 수 있다. 2017년부터 골든글러브 후보 기준이 완화됨으로써, 현재 포지션별 5-6명 이상의 선수가 골든글러브 후보로 선정되고 있다(한국프로야구위원회, 2022)

메이저리그에서는 팀별 감독과 코치 1명씩 투표하고, 미국야구연구협회(SABR)에서 개발한 수비 통계 자료(SDI)를 25% 반영하여 수비 능력이 가장 뛰어난 선수에게 골든글러브를 수상하고 있다. 한국인 선수로는 2012년 추신수 선수가 우익수 부분, 2022년 김하성 선수가 유격수 부분 골든글러브 최종 후보 3인에 선정되었으나, 수상에는 실패하였으나, 2023년 유틸티티 부분에서 김하성 선수가 한국인 최초로 골든글러브를 수상하였다. 한국프로야구 골든글러브 수상자는 프로야구 기자단과 방송 관계자들의 투표로 결정이 되고 있다. 메이저리그처럼 감독 및 코치의 투표, 수비 통계자료를 활용하는 것이 아니라 방송관계자 및 기자단의 투표로 골든글러브 수상자가 결정되다 보니 몇몇의 수상에서 논란이 일어났다. 대표적으로 2012년 A선수-B선수의 골든글러브 투표 논란이 가장 대표적인 사건인데 A선수가 30경기 등판 208이닝 16승 2완투-1완봉승 2.20의 평균자책점을 기록하였고, B선수 27경기 등판 157이닝 17승 0완투-0완봉승 3.59의 평균자책점을 기록하였다. 기록으로 봤을 때 A선수의 수상이 유

력하였으나 B선수가 수상을 하였는데 이는 A선수가 외국인 선수라는 이유와 B선수가 정규리그 우승팀 선수라는 이유가 더해지며 많은 논란이 있었다. 이처럼 한국프로야구 골든글러브 수상에는 방송관계자 및 기자단의 투표로 이루어짐에 따라 포지션별 명확한 기준 및 중요 변인을 알아보는데 한계가 있다.

최근 다양한 스포츠에서 인공지능을 활용한 연구가 많이 이루어지고 있는데 야구에서도 인공지능을 활용한 승패예측, 수상예측과 관련된 다양한 연구들이 이루어지고 있다. 김종훈, 김경태, 한종기(2015)의 Deep Learning 기반 기계학습 알고리즘을 이용한 야구 경기 Big Data 분석 연구에서 30년간의 자료를 분석하여 2015년 시즌의 KBO 야구경기 승패를 예측하는 알고리즘을 제안하였고, 서영진, 문형우, 우용태(2019)의 기계학습 기법을 이용한 한국프로야구 승패 예측 모델의 연구에서는 딥러닝 기법을 활용하여 새로운 형태의 승패 예측 모델을 개발하면서 데이터 구성 및 전처리, 데이터 학습 및 예측, 기대 승률 계산 및 승패 예측 단계로 구분하여 진행하였다. 김태훈, 임성원, 고진광, 이재학(2020)의 인공지능 모델에 따른 한국프로야구 승패 예측 분석에 관한 연구에서는 각 1, 3, 5이닝 별로 가장 정확도가 높으면서 오차가 적은 머신머닝 모델로 KNN과 AdaBoost를 최종 모델로 선정 하여 연구를 진행하였다. 김형우(2021)의 머신러닝 기법을 활용한 프로야구 승패 예측의 연구에서는 로지스틱 회귀분석, 랜덤포레스트, 서포트 벡터 머신 기법으로 세이버 메트릭스를 활용하여 경기 결과를 예측하는 연구가 이루어졌다. 엄대엽, 김성용(2022)의 머신러닝을 이용한 골든글러브 수상 요인 분석에 대한 연구, 엄대엽(2021)의 머신러닝을 이용한 골든글러브 예측모델 개발의 연구에서는 머신러닝 기법 중 하나인 XGBoost를 활용하여 2017~2020년 데이터로 2021년 골든글러브 수상자의 예측비율(%)을 구하여 선수를 예측하는 연구를 진행하였다. 하지만 훈련용 데이터 및 학습용 데이터에 대한 정확한 정의가 없었으며, 짧은 기간(2017~2020)의 데이터를 가지고 수상자의 예측비율(%)만 구하고 비수상자의 예측비율(%)을 제시하지 않았다.

따라서 본 연구에서는 2003년부터 2022년까지 한국프로야구 각 포지션별 골든글러브 후보자 및 수상자의 정량적 범위를 제시한 후 로지스틱 회귀분석과 다양한 머신러닝기법(Support vector machine, 랜덤포레스트, XGBoost)을 학습시켜 각 모델의 성능을 평가한 후 최적화된 인공지능 모델을 탐색하고자 한다.

2. 연구 목적

이 연구는 한국프로야구에서 포지션별 최고의 선수를 뽑는 골든글러브 후보군과 수상자의 기록의 정량적 범위를 탐색한 후 로지스틱 회귀분석 및 머신러닝기법(서포트 백터 머신(Support vector machine), 랜덤포레스트(Random Forest), XGBoost)을 활용하여 각 모델별 투입되는 변수 및 하이퍼 파라미터를 탐색한다. 이후 각 모델의 성능평가를 통해 최적화된 인공지능 모델을 탐색하는데 목적이 있다.

3. 연구 문제

이 연구에서 설정한 연구 문제는 다음과 같다.

문제 1 - 한국프로야구 2003~2022년도 각 포지션별 골든글러브 후보 및 수상자 간 기록의 정량적 범위에는 차이가 있는가?

문제 2 - 한국프로야구 골든글러브 예측을 위한 최적화된 인공지능 모델은 무엇인가?

2-1. 골든글러브 예측 모델 개발을 위해 투입되는 최적화된 변수는 무엇인가?

2-2. 골든글러브 예측 모델 개발을 위한 각 모델별 최적화 알고리즘은 무엇인가?

4. 용어의 정의

1) 한국야구위원회(Korea Baseball organization)

한국야구위원회(KBO)는 영리법인인 구단들을 회원으로 구성된 단체이다(전용배, 2001). 한국프로야구를 관리, 통괄하는 기구로써, 1982년 프로야구 출범(김준한, 2002)과 함께 발족하였고 한국프로야구의 모든 경기(정규리그, 올스타전, 포스트시즌)를 주최한다. 현재 한국야구위원회에 소속된 구단은 삼성 라이온즈, 기아 타이거즈, 두산 베어스, SK 와이번스, KT 위즈, LG 트윈스, NC 다이노스, 한화 이글스, 롯데 자이언츠, 키움 히어로즈 등 10개 구단이다(한국야구위원회, 2020).

2) KBO 골든글러브(KBO Golden Glove Award)

KBO 골든글러브는 KBO 리그에서 매해 각 포지션별로 가장 우수한 활약을 한 선수에게 주어지는 상이다. 예전에는 한국프로야구 골든글러브(Korean Professional Baseball League Golden Glove Award) 였으나, KBO의 브랜드 아이덴디티 통합 작업에 따라 2015년 시즌부터 “KBO 골든글러브” 라는 명칭을 사용하게 되었다.

3) 야구 포지션(Baseball Position)

야구 포지션은 투수(Pitcher), 포수(Catcher), 내야수(Infielder), 외야수(Out fielder), 지명타자(Designated Hitter)로 나눌 수 있다. 투수는 선발 투수, 중간 투수, 마무리 투수로 나눌 수 있으며, 내야수는 1루수, 2루수, 3루수, 유격수, 외야수는 좌익수, 중견수, 우익수로 구분할 수 있다.

4) 정량적 범위(Range of quantitative data)

정량적 범위란 어떤 측정이나 분석에서 사용되는 숫자 또는 수치의 범위를 나타낸다. 어떤 변수나 데이터의 특성을 숫자로 표현하고 어떤 범위 내에 존재하는지를 나타내는 데 도움을 준다. 특정 상황이나 분야에 따라 다를 수 있으며 데이터나 변수의 특성에 따라 정의되기도 한다.

5) 세이버메트릭스(sabermetrics)

세이버 메트릭스란 SABR(The Society for American Baseball Research. 미국야구연구협회) + metrics의 합성어로 야구 기록을 통계학적/수학적으로 분석하는 지표이다. 기존의 관습적 선수 평가론을 부정하고 야구 기록에 대해 좀 더 과학적, 계량적인 평가를 하기 위해 창안된 이론이다.

(1) WAR(Win Above Replacement)

WAR이란 세이버 메트릭스 기록 중 하나로 선수가 팀 승리에 얼마나 공헌하였는가를 종합하여 평가하는 지표이다. 대체 선수 대비 승리기여도로 표현되며 대체 선수는 포지션별 가상의 선수 수준을 의미한다.

(2) wOBA(weighted on-Base Average)

wOBA란 Tom Tango가 창안한 지표로 야구에서 쓰이는 통계 지표로서 타자의 타석당 득점 기여도를 출루율 스케일로 표현한다. 모든 출루를 동등하게 취급하는 출루율과 달리 출루 유형에 따라 가중치를 적용한 지표이다.

$$wOBA = \frac{0.7*(BB-IBB+HBP)+0.9*(1B+ROB)+1.25*2B+1.6*3B+2.0*HR+0.25*SB-0.5*CS}{PA-IBB-SH}$$

수식 1. wOBA를 구하는 계산식

(3) wRAA(Weighted Runs Above Above Replacement)

wRAA란 리그평균대비 득점기여도를 나타내는 지표로서 타자의 wOBA를 누적한 값으로 타석에서 평균보다 얼마나 잘했는지 점수로 환산하였다. wOBA보다 실제 타자의 기여도를 판단하기에 더 적절한 지표이다.

$$wRAA = \frac{wOBA - \lg wOBA}{wOBAscake} * PA$$

수식 2. wRAA를 구하는 계산식

(4) WPA(Win Probability Added)

WPA란 승리 확률기여도를 나타내는 지표로서 두 상황 간의 기대 승률 차이를 나타낸다. 주자 상황, 아웃 개수, 이닝, 점수차의 조합으로 정의되며 기대 승률은 각각의 상황에서 출발했을 때 최종적으로 팀의 승리할 확률을 나타낸다.

(5) wRC(weighted Runs Created)

wRC란 조정득점 창출력을 말한다. 의미 그대로 타자의 득점 생산력을 나타내며, 현존하는 타격 스탯 중 가장 정확한 타격 스탯으로 각광받고 있다.

$$wRC = (\frac{wOBA - \leq ague\, wOBA}{wOBA}) \bullet PA$$

수식 3 . wRC를 구하는 계산식

(6) IsoP(Isolated Power)

IsoP란 타자의 순수한 장타율을 나타낸 지표로서 장타율에 1루타가 포함되는 단점을 보완하기 위해 만들어진 지표이다.

IsoP = (장타율 - 타율)

(7) IsoD(Isolated Disciple)

IsoD란 순수 출루율을 나타내는 지표로서 볼넷, 고의4구등 선수의 선구안을 알아볼 수 있는 지표이다.

IsoD = (출루율 - 타율)

(8) BABIP(Batting Average on Balls In Play)

BABIP란 인플레이 타구의 안타비율을 나타내는 기록 중 하나이다. 경기 상황 중 타구가 경기장 내에 머문 타구에 대한 타율이며 홈런은 BABIP에서 제외된다.

$$BABIP = \frac{H - HR}{AB - K - HR + SF}$$

수식 4. BABIP를 구하는 계산식

(9) WHIP(Walks Plus Hits Divided by innings Pitched)

WHIP란 야구에서 투수의 성적을 평가하는 지표 중 하나로서 이닝당 출루허용률을 나타내는 지표이다. 피안타 수와 볼넷 수의 합을 이닝으로 나누는 지표로서 몸에 맞는볼을 포함하지 않는다.

WHIP = (피안타+볼넷) / 이닝

(10) FIP(Fielding Independent Pitching)

FIP란 수비무관 평균자책점으로 수비와 무관한 홈런, 삼진, 볼넷, 몸에맞는 볼의 4가지 기록만을 가지고 평균자책점을 도출하는 지표이다

$$FIP = [\frac{(HR*13) + (3*(\text{볼넷} + \text{몸에 맞는 볼})) - (2*\text{삼진})}{\text{이닝}}] + FIP\text{상수값}$$

수식 5. FIP를 구하는 계산식

Ⅱ. 이론적 배경

1. 한국 프로야구 골든글러브

1) 골든글러브 선정

연도의 수비, 공격, 인기도를 종합한 수상자를 투표인단이 선정하며 (2023 KBO 리그 규정, 2023), 선정방법은 KBO에서 매년 취재기자, 중계 PD, 해설위원 등 미디어 관계자들을 대상으로 투표를 진행하고 최다 투표를 받은 포지션별 최고의 선수에게 골든글러브상을 수여한다(엄대엽, 2022).

2) 연도별 골든글러브 선정 조건의 변화

프로야구 원년 시즌이었던 1982년과 1983년 2년간은 골든글러브가 MLB나 NPB처럼 공격이 아닌 수비율로 시상되었고, 1984년부터는 이 두 부문이 통합되었으며, 대체로 타격 지표를 우선하여 시상한다. 구체적인 기준은 타고투저나 투고타저 등 각 시즌의 성향에 따라 조금씩 변화되었다가 형평성에 문제가 있어 후보 기준을 대폭 완화하였다. 2003년부터 2016년까지는 야수는 시즌에 따라 차이는 존재하지만 수비 출전 경기수, 규정타석 이상, 포지션별로 타율도 어느 수준 이상이 후보가 될 수 있었다. 투수 또한 평균자책점 3점대 이하, 14~15승 이상, 30세이브 이상이라는 규정이 다소 빡빡하게 존재해서 한 시즌에 야수 포지션은 2 ~ 5명, 투수 포지션은 4 ~ 7명 정도의 후보가 배출되었다. 하지만 2017년 기준이 대폭 완화되었다. 투수 포지션에서는 10승 또는 30세이브 이상, 30홀드 이상의 선수가 골든글러브 후보가 되었으며, 야수부분에서는 내야수 및 외야수 수비 출전 팀 경기수 X 5이닝(720이닝) 이상, 지명타자는 규정타석의 2/3으로 297타석 이상을 출전하면 후보가 된다. 따라서 2017년도부터 투수는 20명이상의 선수가 골든글러브 후보가 되었으며, 각 포지션별로 5 ~ 9명의 선수가 골든글러브 후보가 되었다.

3) 미국프로야구 골든글러브와의 차이점

미국프로야구(MLB)의 경우 한국 프로야구(KBO)와 다르게 각 포지션별 최고의 수비력을 보여준 선수를 뽑는 “골든글러브”와 각 포지션별 최고의 공격력을 보여준 선수를 뽑는 “실버슬러거(Silver Slugger Award)”상으로 구분되어있다. 미국프로야구의 골든 글러브 선정은 30개 구단 감독과 6명의 코치진의 투표가 75%, 미국 야구연구 협회의 수비지표 25%를 합산하여 수상자를 결정하고 있으며, 골든글러브는 수비능력만을 고려하여 수상자를 뽑고, 실버슬러거의 경우 타격 능력만을 고려하여 수상자를 뽑고 있다. 미국프로야구의 골든글러브는 기자들이 뽑는 MVP나 사이 영 상과는 다르게 감독과 코치들이 뽑는다. 각 팀마다 감독과 코치 1명(팀당 2명)을 선정하여 투표하는데, 자신이 속한 팀에 속한 선수는 뽑을 수 없다. 그래서인지 그해 상대했던 몇 경기의 기억과 이전의 이미지로 뽑는다는 비난이 심하다. 심지어 1999년 1루수로 28경기만 뛰고 지명타자로 135경기를 뛴 선수가 1루 골든글러브를 수상한 것 등 실제 경기를 뛴 포지션과 다르게 선정되는 사례들이 있어 골든글러브 수상에 대한 의문이 제기되었다. 그래서 2013년부터 세이버메트릭스 수치 중 수비 관련 통계 자료(SDI)를 투표에 반영하기로 하였다. 미국야구연구협회(SABR)에서 투표인단인 각 팀 감독과 코치들에게 SDI를 배포하기도 했다. 실제 SDI가 투표에 반영된 비율은 약 25%라고 한다. 2013년 골든글러브 수상자들의 발표 이후 팬들은 기존보다 대체적으로 납득한다는 평이지만 평균 이하의 수비수가 세이버 반영을 한 이후에도 수상한 사례들이 있어 여전히 신뢰성에 의문부호가 붙었다. 한국프로야구에서 ‘실버슬러거’와 같은 공격 능력을 고려한 수상이 없다 보니 ‘골든글러브’의 수상자의 수비능력보다 각 포지션의 최고의 공격 지표를 기록한 선수 혹은 최고의 인기를 가진 선수들이 골든글러브를 받고 있다.

2. 정량적 자료분석의 중요성

1) 세이버 메트릭스의 정의

세이버 메트릭스는 야구에 사회과학의 게임 이론과 통계학적 방법론을 적극 도입하여 기존 야구 기록의 부족한 부분을 개선하고, 선수의 가치를 비롯한 '야구의 본질'에 대해 학문적이고 깊이 있는 접근 하는 것을 가리킨다. 기존의 관습적 선수 평가론을 부정하고, 야구선수에 대해 좀 더 과학적, 통계적인 평가를 하기 위해 창안된 이론이다. 간단하게 정의하면, 누적된 야구 기록을 활용하여 통계학적으로 분석을 하는 분야를 세이버메트릭스라고 하며, 또한 세이버메트릭스 자료를 분석하는 사람을 세이버메트리션이라고 부른다(홍종선 등, 2016).

미국야구연구학회(SABR: Society for American Baseball Research)의 빌 제임스(Bill James)가 SABR과 metrics의 합성어인 세이버 메트릭스를 처음 사용하였다(SABR, 2016). 세이버 메트릭스는 야구 기록을 통계학적/수학적으로 분석한 자료로 많은 기록 자료가 만들어졌다 (안현호, 2015). 타격지표로는 OPS(on-base prercen- tage Plus slugging percentage, 출루율+장타율), BABIP(Batting Average on Ball In Play, 인플레이 타구에 대한 타율), RC(Run Created, 득점창출 능력), IsoP(Isolated Power, 순수장타율)등이 대표적 이며, 투수지표로는 WHIP(Walk and Hit per Innings Piched, 이닝당 출루 허용율), K/BB(삼진하나당 볼 넷비율), FIP(Fielding Independent Pitching, 수비 무관 평균자책점), Rel%(승계주자 실점율)등이 대 중적으로 알려져 있다. 또한 타격능력, 수비능력, 투구능력을 모두 합쳐 통계적으로 나타낸 수치인 WAR(Wins Above Replacement,대체선수 대비 승리 기여도)가 가장 대표적인 세이버 메트릭스 지표 라 할 수 있다(권순규 등, 2019).

2) 스포츠 경기분석의 통계적 접근

(1) 스포츠경기분석

스포츠경기분석을 영어로 직역하면, Game analysis of sports 또는 Match analysis of sports 정도가 된다. 반면에 스포츠경기분석의 실질적인 영역은 다양한 용어와 함께 그 범위가 보다 융합적이다. 스포츠경기분석의 실질적인 영역은 스포츠 매치 분석, 스포츠 게임 분석, 스포츠 움직임 분석, 스포츠 수학적 분석, 스포츠 통계 분석, 스포츠 동작 분석, 스포츠 전력 분석 등 다양한 용어로 표현되며, 그 범위는 융합적이다(McGarry, O’ Donoghue, & Sampaio, 2013). 스포츠 경기에서 나타나는 경기력을 기술하기 위해, 경기 중 발생하는 주요 사건이나 내용을 체계적으로 관찰하고 기록하는 데 중점을 둔다(최형준, 2022). 이는 선수와 팀의 성능 향상을 도모하고, 전략적인 게임 플레이를 구축하는 데 있어 필수적인 요소이다.

(2) 데이터 수집 및 전처리

스포츠 경기 분석은 선수들의 성능을 평가하고 경기 결과를 예측하는 데 사용되어 왔으며, 통계학은 이 과정에서 중요한 역할을 담당하고 있다(Hughes, & Franks, 2004). 경기 데이터의 수집은 센서 기술, 비디오 분석, 수동 입력을 통해 이루어지며, 이후 데이터 전처리 과정에서는 불완전하거나 잘못된 데이터를 정제하고 분석에 적합한 형태로 변환하는 작업이 수행된다. 이러한 과정을 통해 얻어진 데이터는 다음 단계의 통계적 분석을 위한 기반이 되며, 정확한 분석 결과를 도출하기 위해서는 데이터의 질이 매우 중요하다.

(3) 통계적 접근법

스포츠 데이터에는 다양한 통계적 방법론이 적용될 수 있으며, 이러한 방법론을 통해 선수의 성능을 평가하고 경기의 결과를 예측하며 최적의 전략을 개발할 수 있다(James, Witten, Hastie, & Tibshirani, 2013). 통계적 방법론의 적용은 스포츠 분석을

경기 성과 향상, 부상 예방, 선수 선발, 팬 경험 향상 등의 다양한 분야로 확장할 수 있게 해준다(Pappalardo et al, 2019). 이를 통해 스포츠 팀과 관계자들은 더 정보에 기반한 의사 결정을 할 수 있게 되며, 스포츠 경기 분석과 통계의 연관성은 계속해서 발전해 나가고 있다.

3. 머신러닝 기법

1) 머신러닝의 정의

머신러닝은 데이터를 기반으로 패턴을 학습하고 이를 바탕으로 예측, 분류, 군집화 등의 다양한 작업을 수행하는 알고리즘과 모델을 연구하는 분야이다. 이는 컴퓨터가 명시적으로 프로그래밍 되지 않아도 데이터로부터 학습할 수 있는 능력을 포함한다(Bishop, & Nasrabadi, 2006). Arthur Samuel은 1959년에 머신러닝을 “컴퓨터가 명시적으로 프로그래밍 되지 않아도 학습할 수 있는 능력“이라고 정의하며 이 분야의 개척자 중 한 명으로 꼽힌다(Samuel, 2000).

2) 머신러닝의 종류

머신 러닝은 학습의 구분에 따라서 구분한다. 먼저, 지도학습 (Supervised Learning): 입력과 출력 데이터의 예시를 바탕으로 함수적 관계를 학습하는 방법을 말한다. 지도학습을 통하여. 분류(classification)와 회귀(regression) 문제를 풀이하는데 사용된다(Hart, Stork, & Duda, 2000). 지도학습에 반대되는 개념이 비지도학습 (Unsupervised Learning)이다. 비지도학습이란 출력 데이터 없이 입력 데이터의 패턴이나 구조를 찾는 학습 방식으로, 군집화(clustering)와 차원 축소(dimensionality reduction)가 대표적인 예시이다(Hinton, & Salakhutdinov 2006). 또한 환경과의 상호작용을 통해 얻은 보상을 바탕으로 최적의 행동을 학습하는 방식을 강화학습 (Reinforcement Learning)이라고 한다. 이는 특히 게임이나 로봇 제어 같은 분야에서 사용된다(Sutton, & Barto, 2018). 준지도 학습 (Semi-supervised Learning)은 레이블이 있는 데이터와 없는 데이터를 모두 사용하며, 레이블링 비용이 높거나 어려운 경우에 유용하다(Chapelle et al, 2006). 자기지도 학습 (Self-supervised Learning)은 레이블이 필요 없는 지도 학습의 한 형태로, 데이터 자체에서 레이블을 생성한다. 예를 들어, 이미지 처리에서 일부를 가리고 해당 부분을 예측하게 하는 방식이 있다(Goodfellow et al., 2016). 마지막으로 전이 학습 (Transfer Learning)은 한 분야에서 학습된 모델을 다른 분야에 적용하며,

데이터가 제한적인 상황에서 유용하다(Pan & Yang, 2010).

3) 학습의 원리

(1) 지도학습 (Supervised Learning)

- 데이터 및 레이블 : 지도학습에서는 학습 노델에게 입력 데이터와 그에 상응하는 정답레이블을 제공한다.
- 학습 알고리즘 : 모델은 입력 데이터와 정답 레이블 간의 관계를 학습하기 위해 알고리즘을 사용한다. 대표적인 알고리즘으로는 선형 회귀, 결정트리, 신경망 등이 있다.
- 손실 함수 및 최적화 : 모델은 예측과 실제 레이블 간의 차이를 나타내는 손실 함수를 최소화하도록 학습된다. 최적화 알고리즘을 사용하여 모델의 가중치 및 편향을 조정하여 손실을 최소화한다.
- 예측 : 학습된 모델은 새로운 입력에 대한 예측을 수행할 수 있다.

(2) 비지도학습 (Unsupervised Learning)

- 데이터만 주어짐 : 비지도학습에서는 레이블이 없는 데이터만 주어진다
- 패턴 발견 및 군집화 : 모델은 데이터 내의 패턴이나 구조를 발견하거나 비슷한 특성을 갖는 데이터를 군집화한다. 대표적인 알고리즘으로는 K-평균 군집화, 계층적 군집화, 차원 축소 기법 등이 있다

(3) 강화학습 (Reinforcement Learning)

- 에이전트와 환경 : 강화학습에서는 에이전트가 주어진 환경에서 행동하고 그 결과로 보상 또는 패널티를 받는다.
- 학습과정 : 에이전트는 환경과의 상호작용을 통해 보상을 최대화하는 최적의 정책(Policy)을 학습한다. 이는 시행착오를 통해 이루어진다.
- 탐험 및 이용 : 에이전트는 탐험(exploration)과 이용(exploitation)을 균형있게

수행하여 미래의 보상을 최대화하려고 노력한다.

- 머신러닝은 데이터 기반으로 모델을 훈련시키는 과정이므로 데이터의 품질과 다양성이 중요하다. 또한 모델의 성능을 평가하고 개선하기 위한 지속적인 과정이 필요하며, 이를 위해 교차검증, 하이퍼파라미터 튜닝등의 기법이 사용된다.

4) 머신러닝과 수학

머신러닝은 선형대수학, 확률론, 통계학, 최적화 이론 등 다양한 수학적 개념을 기반으로 한다. 이러한 수학적 도구들은 머신러닝 모델이 데이터에서 복잡한 패턴을 학습하고 예측을 수행할 수 있게 한다(Hastie, Tibshirani, & Friedman, 2009).

따라서 머신러닝은 데이터로부터 지식을 추출하고 예측 모델을 구축하는데 사용되는 강력한 도구로, 그 적용 분야는 의학, 금융, 교육, 산업 등 매우 광범위하다. 이러한 다양한 분야에서 머신러닝의 알고리즘과 모델들이 활용되며, 그 중요성은 계속해서 증가하고 있다. 머신러닝의 지속적인 연구와 발전은 더욱 정확하고 효율적인 예측 모델을 만들어내며, 현대 기술의 발전을 이끌고 있다.

4. 로지스틱 회귀분석

로지스틱 회귀분석은 이진 또는 다항 분류 문제에 널리 사용되는 통계적 방법이며, 복잡한 머신러닝 모델에 비해 해석이 용이한 장점을 지닌다.

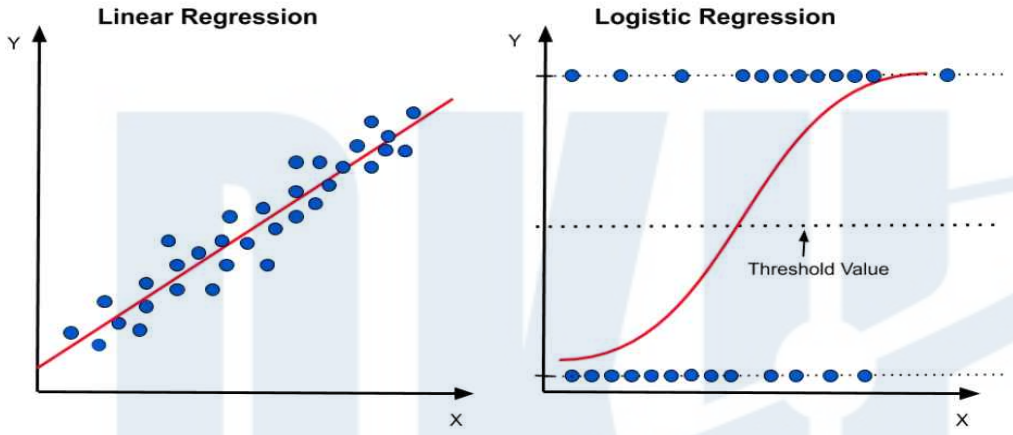


그림 1. 로지스틱 회귀분석 모델

(그림출처 : https://velog.io/@73syjs/Logistic-Regression)

1) 로지스틱 회귀분석의 개요

로지스틱 회귀분석(logistic regression)은 영국의 통계학자인 D.R.Cox가 1958년에 제안한 확률 모델로서 독립변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계기법이다. 로지스틱 회귀는 선형 회귀분석과는 다르게 종속변수가 범주형 데이터를 대상으로 하며, 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류기법으로 볼 수 있다. 흔히 로지스틱 회귀는 종속변수가 이항형 문제를 지칭할 때 사용된다. 이외에 두 개 이상의 범주를 가지는 문제가 대상인 경우엔 다항 로지스틱 회귀(multinomial logistic regression) 또는 분화 로지스틱회귀(polytomous logistic regression)라고 하고 복수의 범주이면서 순서가 존재

하면 서수 로지스틱 회귀(ordinal logistic regression)라고 한다. 로지스틱 회귀분석은 결과 변수가 범주형 데이터인 경우, 특히 이진 분류 문제에 주로 사용된며(King, & Zeng, 2001), 이 방법은 입력 변수와 결과 변수 간의 관계를 모델링하며, 로지스틱 함수를 사용하여 확률을 예측한다.

2) 로지스틱 함수와 확률 예측

로지스틱 회귀모델은 선형 회귀모델의 출력을 로지스틱 함수를 통해 변환하여 확률을 예측한다(Cessie, & Houwelingen, 1992). 로지스틱 함수는 S자 형태를 가지며, 출력 값의 범위를 [0, 1]로 제한한다.

3) 모델 추정과 최적화

로지스틱 회귀모델의 계수는 주로 최대우도 추정법을 사용하여 추정되며, 이 과정에서 로그 손실 함수를 최소화하는 것이 목표이다. 최근 연구에서는 기존의 최적화 방법 외에도 다양한 정규화 기법과 최적화 알고리즘을 적용하여 모델의 성능을 향상시키고 있다(Friedman, Hastie, & Tibshirani, 2010).

4) 모델 평가와 해석

로지스틱 회귀모델의 성능 평가는 주로 ROC 곡선 및 AUC 값을 통해 이루어진다(Bradley, 1997). 또한, 계수의 크기와 부호를 통해 각 변수가 결과 변수에 미치는 영향을 해석할 수 있다. 로지스틱 회귀분석은 그 해석 가능성과 실용성으로 인해 여전히 활발히 사용되는 모델 중 하나이다. 최근의 연구 동향을 반영한 로지스틱 회귀분석의 이해는 데이터 과학과 머신러닝 분야에서 중요한 역할을 한다.

5. 서포트 벡터 머신(Support vector machine)

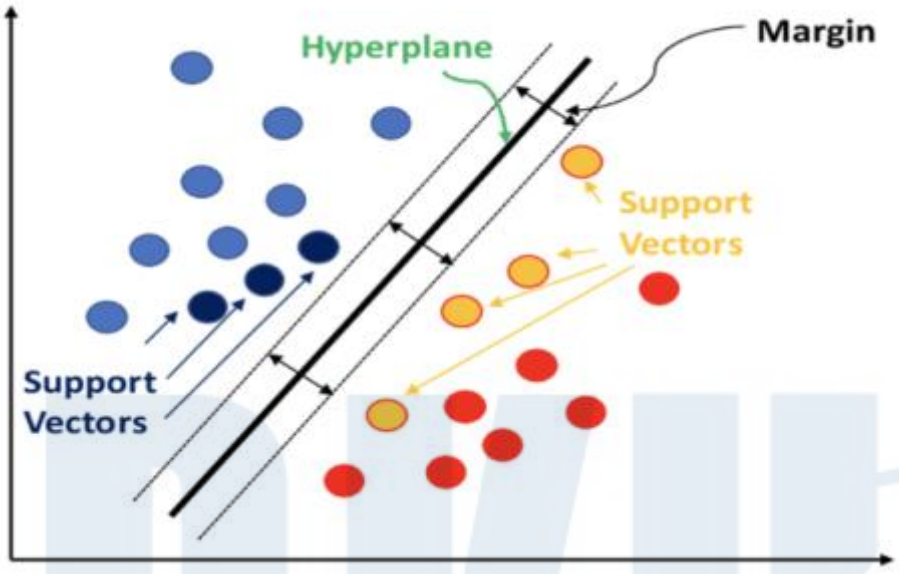


그림 2. Support vector machine 모델

(그림 출처 : https://datatron.com/what-is-a-support-vector-machine/)

서포트 백터 머신(Support Vector Machine)은 분류 문제를 해결하기 위해 설계된 강력한 머신러닝 알고리즘이다. 본 논문에서는 SVM의 이론적 배경, 기본 원리, 그리고 수학적 배경에 대해 심층적으로 분석하며, 특히 결정 경계와 마진 최적화의 중요성을 강조한다. SVM의 주된 목표는 두 클래스 사이의 마진을 최대화하여 데이터를 효과적으로 분류하는 최적의 초평면을 찾는 것이다(Cortes, & Vapnik, 1995).

SVM 알고리즘은 두 클래스의 데이터를 분류하기 위해 결정 경계를 사용한다. 결정 경계는 초평면의 형태를 취하며, 서포트 벡터라 불리는 특정 데이터 포인트들에 의해 결정된다. 이 서포트 벡터들은 결정 경계에 가장 가까운 포인트들이며, 마진은 이 서포트 벡터들과 결정 경계 사이의 거리로 정의된다. SVM 알고리즘의 핵심 목적은 이 마진을 최대화하는 것이다(Vapnik, 1999).

SVM의 수학적 배경은 초평면의 수식화와 관련된 선형 대수 및 최적화 개념에 깊이

뿌리를 두고 있다. 초평면은 $w \cdot x+b=0$ 형태의 방정식으로 표현될 수 있으며, 여기서 w는 가중치 벡터이고, x는 특성 벡터, b는 편향이다. SVM 알고리즘은 이 초평면을 조정하여 마진을 최대화하는 방향으로 학습한다(Cristianini, & Shawe-Taylor, 2000).

그러나 많은 실제 데이터셋은 선형적으로 구분될 수 없는 복잡한 구조를 가지고 있다. SVM은 이러한 비선형 데이터셋에도 적용될 수 있도록 커널 트릭을 사용한다. 커널 트릭은 데이터를 더 높은 차원의 공간으로 변환하여 선형적으로 구분할 수 있게 해준다. 대표적인 커널 함수로는 다항 커널, RBF (Radial Basis Function) 커널, 시그모이드 커널 등이 있다(Boser, Guyon, & Vapnik, 1992). 이러한 기술을 통해 SVM은 다양한 유형의 데이터에 대해 강력하고 유연한 성능을 발휘한다.

결론적으로, SVM은 마진을 최대화하는 결정 경계를 찾아내어 데이터를 정확하게 분류하는 효과적인 방법을 제공한다. 커널 트릭을 활용하여 비선형 데이터에도 적용 가능하며, 그 이론적 배경은 선형 대수와 최적화 이론에 기반을 두고 있다.

6. 랜덤포레스트(Random Forest)

본 논문에서는 앙상블 학습 방법의 하나인 랜덤포레스트(Random Forest) 알고리즘의 이론적 배경과 주요 원리에 대해 다룬다. 랜덤포레스트는 다수의 결정 트리를 결합하여 분류와 회귀 문제에 모두 사용될 수 있는 강력한 머신러닝 알고리즘이다. 이 알고리즘은 각 결정 트리가 독립적으로 훈련되며, 다양한 트리들을 통해 최종 예측을 수행하는 방식으로 작동한다. 이러한 방식은 모델의 성능을 향상시키고 과적합을 방지하는 데 도움을 준다(Breiman, 2001).

랜덤포레스트의 핵심은 두 가지 무작위화 전략, 즉 부트스트랩 샘플링과 특성의 랜덤 선택에 있다. 부트스트랩 샘플링은 각 결정 트리를 원본 훈련 데이터의 재샘플링된 버전으로 훈련시키며, 특성의 랜덤 선택은 각 노드에서 최적의 분할을 결정할 때 모든 특성을 고려하는 대신 무작위로 선택된 일부 특성만을 고려하도록 한다(Fron, 1992). 이러한 전략들은 랜덤포레스트의 각 결정 트리가 서로 다르게 만들어지도록 하여 모델의 다양성을 증가시키고 과적합을 줄인다.

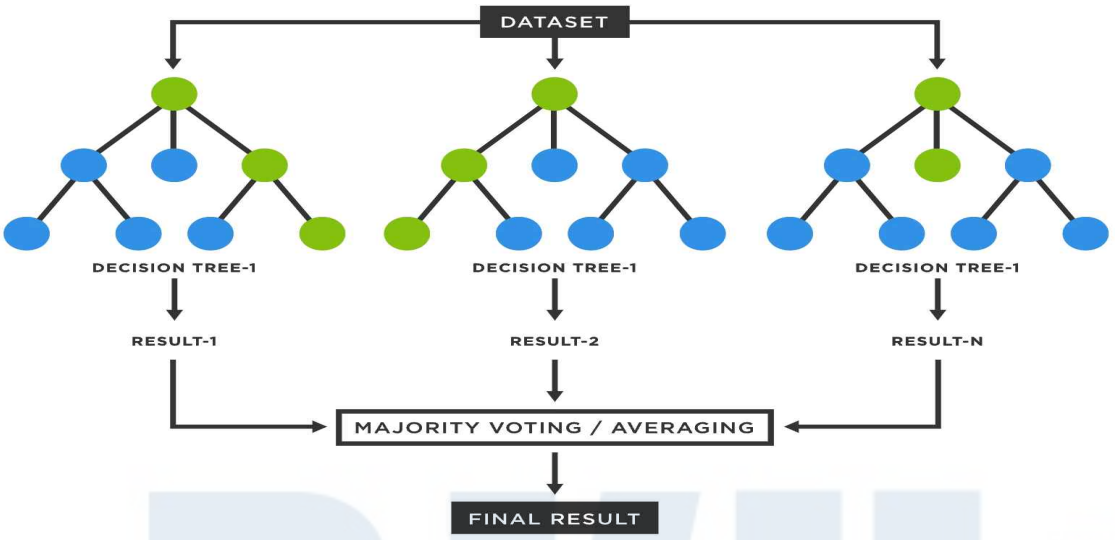


그림 3. 랜덤포레스트 모델

(Nguyen, Wang & Nguyen (2013))

분류 문제에서 랜덤포레스트의 예측은 개별 결정 트리의 예측을 종합하여 가장 많은 표를 얻은 클래스를 최종 예측값으로 선택한다. 반면, 회귀 문제에서는 개별 결정 트리의 예측값의 평균을 사용하여 최종 예측값을 결정한다(Liaw & Wiener, 2002; Loh, 2011). 이러한 방식은 랜덤포레스트를 다양한 유형의 데이터에 대해 유연하게 적용할 수 있게 한다.

결론적으로, 랜덤포레스트는 높은 예측 성능, 특성 중요도 평가, 과적합 방지 등의 장점을 지닌 강력한 머신러닝 알고리즘이다. 그러나 모델의 해석이 어렵고 훈련 시간이 길 수 있다는 단점도 존재한다.

7. XGBoost

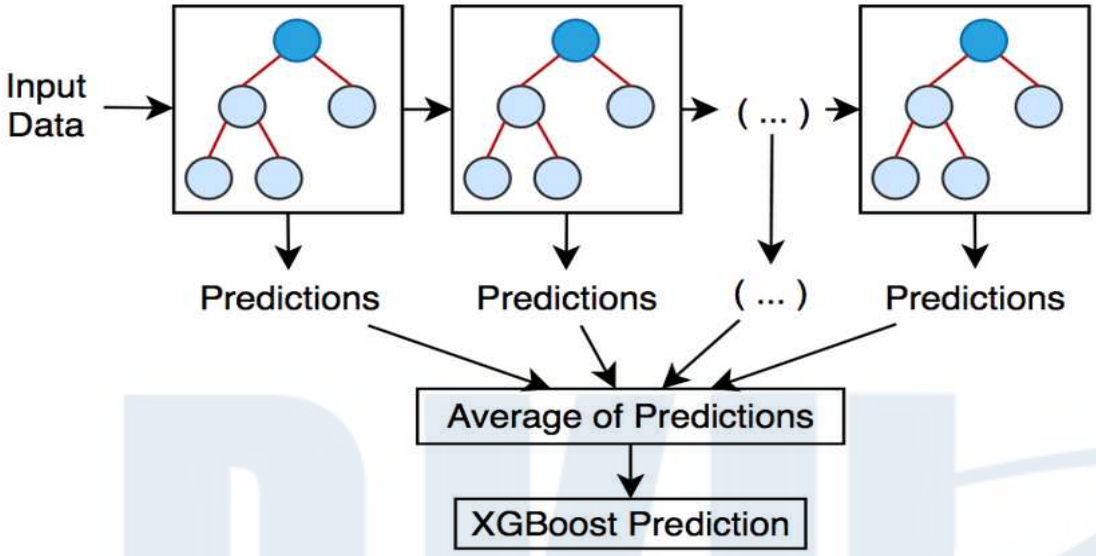


그림 4. XGBoost 모델

(Chen, & Guestrin, (2016))

XGBoost는 고성능과 확장성을 갖춘 트리 부스팅 시스템으로, 빅데이터 환경에서도 효과적으로 작동하며 분류와 회귀 문제에서 뛰어난 성능을 보인다. 본 알고리즘은 그래디언트 부스팅을 기반으로 하며, 정규화, 병렬 처리, 자동 가지치기, 내장 교차 검증 등 다양한 기술적 개선을 통해 기존 방법론들을 뛰어넘는 성능을 제공한다(Friedman, 2001).

그래디언트 부스팅은 약한 학습기들을 순차적으로 학습시켜 강한 학습기를 만드는 앙상블 기법이다. 이 과정에서 각 단계에서 발생하는 오차를 줄이는 방향으로 모델을 점진적으로 개선하여 성능을 향상시킨다(Chen, & Guestrin, 2016). XGBoost는 이러한 그래디언트 부스팅 방법론을 확장하여, 모델의 정규화와 효율적인 학습 과정을 통해 뛰어난 성능을 달성한다. 또한, 병렬 처리 기능을 통해 노드 생성 과정을 빠르게 처리하여 전체 학습 시간을 단축시킨다(Chen, & Guestrin, 2016).

XGBoost의 목적 함수는 손실 함수와 정규화 항으로 구성된다. 손실 함수는 모델의

예측 오차를 측정하고, 정규화 항은 모델의 복잡성에 페널티를 부여하여 과적합을 방지한다. XGBoost는 이 목적 함수를 최적화하여 모델의 성능을 향상시킨다(Chen et al,, 2015). 이 과정에서 내장된 교차 검증 기능을 통해 모델의 성능을 지속적으로 평가하고, 최적의 반복 횟수를 찾아낸다(Chen et al, 2015).

결론적으로, XGBoost는 그래디언트 부스팅 기법의 확장 버전으로, 정규화, 병렬 처리, 자동 가지치기, 내장 교차 검증 등의 특징을 통해 대규모 데이터셋 처리에 있어 뛰어난 성능과 효율성을 제공한다.

Ⅲ. 연구방법

이 연구는 한국 프로야구에서 포지션별 최고의 선수를 상징하는 “골든글러브” 상의 후보 및 수상자를 조사하여 정량적 범위를 나타내고, 더 나아가 선수들의 기록을 토대로 인공지능 기법을 활용하여 예측모델을 개발 및 비교하는데 있다.

1. 연구대상

이 연구의 대상은 2003년부터 2022년까지 한국프로야구 골든글러브 후보 및 수상자의 기록을 조사하였다. 20년간의 선수 기록으로 투수(264명), 타자(858명) {포수(89명), 1루수(99명), 2루수(93명), 3루수(98명), 유격수(98명), 외야수(296명), 지명타자(85명)}으로 수집하였다. <표 1>은 2003년부터 2022년까지 포지션별 골든글러브 후보 선수의 수를 정리한 것이다.

표 1. 2003년 ~ 2022년 골든글러브 후보 수

구분	선수 수
투수	264명
포수	89명
1루수	99명
2루수	93명
3루수	98명
유격수	98명
외야수	296명
지명타자	85명
전체	1,122명

<표 2>는 프로야구 타자기록을 설명하는 표이다. 기본 기록과 세이버 메트릭스 기록을 구분하였다.

표 2. 프로야구 타자기록(기본 기록, 세이버메트릭스 기록)

구분	종류
타자 기본 기록	게임 수, 타석 수, 타수, 득점 수, 안타 수, 2루타 수, 3루타 수, 홈런 수, 루타 수, 타점 수, 도루 수, 도실 수, 볼넷 수, 사구 수, 고의4구 수, 삼진 수, 병살 수, 희타 수, 희비 수, 타율, 출루율, 장타율, HR%, BB%, K%, BB/K, spd, PSN
타자 세이버메트릭스 기록	OPS, wOBA, wRC. IsoP, IsoD, BABIP, WAR, wRAA, WPA

<표 3>은 프로야구 투수기록을 설명하는 표이다. 기본 기록과 세이버 메트릭스 기록을 구분하였다.

표 3. 프로야구 투수기록(기본 기록, 세이버메트릭스 기록)

구분	종류
투수 기본 기록	출장 수, 완투 수, 완봉 수, 승리 수, 패배 수, 세이브 수, 홀드 수, 이닝 수, 실점 수, 자책점 수, 상대타자 수, 피안타 수, 피홈런 수, 볼넷 수, 고의4구 수, 사구 수, 삼진 수, ERA, 보크 수, 폭투 수, K/9, BB/9, K/BB, HR/9, K%, BB%, K-BB%, IP/G, P/G, P/IP, P/PA상대타자 타율, 상대타자 출루율, CYP, 투구 수
투수 세이버메트릭스 기록	FIP, WHIP, BABIP, LOB%, PFR, WAR, WPA

2. 자료수집 도구

이 연구는 인터넷을 통한 자료 조사를 실시하였다. 한국프로야구 골든글러브 후보 및 수상자를 알아보기 위하여 한국야구위원회(www.koreabaseball.com)에서 검색하여 자료수집 하였다. <그림 5>는 KBO 보도자료 중 골든글러브 후보 선수의 기록에 대한 보도자료이다. 골든글러브 후보 및 수상 선수의 기록을 수집하기 위하여 한국프로야구 통계사이트인 스텟티즈(www.statiz.com)와 한국야구위원회(www.koreabaseball.com)에서 자료를 수집하였다.

KBO 보도자료

제목 | 골든글러브 후보 | 검색

번호	제목	첨부	등록일	조회수
20	2022 신한은행 SOL KBO 골든글러브 후보 확정		2022.11.28	17441
19	2021 신한은행 SOL KBO 골든글러브 후보 확정		2021.12.01	18639
18	2020 신한은행 SOL KBO 골든글러브 후보 확정		2020.12.02	17095
17	2018 신한은행 MY CAR KBO 골든글러브 후보 확정		2018.12.03	33872
16	2017 타이어뱅크 KBO 골든글러브 후보 확정		2017.12.04	31608
15	2016 타이어뱅크 KBO 골든글러브 후보 확정		2016.12.05	3419
14	2015 타이어뱅크 KBO 골든글러브 후보 확정		2015.11.30	9654
13	2014 프로야구 골든글러브 후보 확정		2014.11.30	5361
12	2013 프로야구 골든글러브 후보 확정		2013.11.27	11830

그림 5. 골든글러브 후보 검색 엔진

<그림 6>은 MS Excel에 수집된 각 포지션별 후보 및 수상 선수의 기록을 정리한 예시이다.

2020

이름	팀	WAR*	G	타석	타수	득점	안타	2타	3타	홈런
박민우	NC	4.48	126	530	467	82	161	27	5	8
최주환	두산	4	140	573	508	63	155	29	4	16
박경수	KT	2.39	119	391	324	33	91	17	0	13
정주현	LG	0.03	134	371	328	50	81	10	4	4
안치홍	롯데	2.01	124	460	412	49	118	28	0	8
김상수	삼성	3.27	120	471	404	71	123	18	3	5

2021

이름	팀	WAR*	G	타석	타수	득점	안타	2타	3타	홈런
김상수	삼성	0.38	132	496	429	46	101	17	1	3
서건창	LG	2.83	144	599	512	78	130	24	2	6
안치홍	롯데	3.45	119	490	421	58	129	30	2	10
김선빈	KIA	3.42	130	564	501	55	154	32	0	5
정은원	한화	4.47	139	608	495	85	140	22	5	6

2022

이름	팀	WAR*	G	타석	타수	득점	안타	2타	3타	홈런
김혜성	키움	4.8	129	566	516	81	164	18	7	4
김선빈	KIA	2.88	140	587	505	51	145	23	0	3
박민우	NC	2.13	134	487	444	54	117	28	1	10
김지찬	삼성	2.08	113	429	361	62	101	7	6	0
안치홍	롯데	3.24	132	562	493	71	140	27	3	14
강승호	두산	2.34	134	487	444	54	117	28	1	10
정은원	한화	3.7	140	601	508	67	140	20	2	8

그림 6. 골든글러브 후보 및 수상선수 기록 정리 예시

3. 연구절차

이 연구는 2003~2022년 한국프로야구 골든글러브 후보 및 수상자의 기록을 대상으로 머신러닝 모델을 학습시켜 최적의 모델을 만들고자 연구를 설계하였다. 전 포지션(투수, 포수, 1루수, 2루수, 3루수, 유격수, 외야수, 지명타자)를 대상으로 자료수집을 실시하였으며, 파이썬(Python 3.10.11) 프로그램을 이용하였다. 각 포지션별 중요 변인을 추출하였으며, 각 모델에 맞는 하이퍼 파라미터를 적용하여 예측모델을 설계하였다. F1 Score와 정확도를 토대로 최적의 모델을 선정하였다. 골든글러브 수상자 예측을 위해 로지스틱 회귀분석과 머신러닝 기법의 예측모델을 설계하였으며, <표 4>와 같이 도식화하였다.

표 4. 골든글러브 수상자 예측 모델 개발 연구 절차

준비	연구 설계 ㅣ 자료 수집 ㅣ (대상)(row/column)
자료 전처리 및 변수선택	**1. 자료 전처리** • 데이터정리 • 결측치처리 : 제외, 0값, 평균, 분산 • 인코딩(명목-원핫인코딩,연속변수 : 최대-최소 정규화) **2. 변수선택** • 로지스틱 회귀 (변수 설명도 높은 변수 선택) • 모델링을 위한 변수 최종 선택 • 포지션별 최종 데이터 행렬 크기 (row/ column)
모델 개발	**1. 개발 환경** • 파이썬 언어 • 패키지 **2. 변수 구성** • 모델링 대상 : 03-22시즌 데이터 • 독립 : 0000… • 종속 : 골든글러브 수상여부 [1:수상], [0:비수상] **3. 알고리즘 개발** • 알고리즘 개발 및 비교 • 인공신경망, XGBoost, SVM, 랜덤포레스트 • 각 알고리즘 설계 (하이퍼파라미터) **4. 성능평가** • 알고리즘 별 개발 비교(정확도, 정밀도, 재현율, F1score) • 변수중요도,
적용	**1. 모델 선택** • 최적 모델 선택

4. 자료처리

이 연구의 자료처리 방법으로 SPSS 25.0버전을 사용하여 기술통계를 실시하였다. 또한 머신러닝 기법의 골든글러브 수상자 예측 모델을 개발하기 위하여 파이썬(Python)프로그램을 이용하였으며 데이터 전처리, 모델선정, 성능 평가의 과정을 거쳐 인공지능 예측 모델을 개발하였다.

1) 데이터 전처리

인공지능 모델 개발을 위해 수집된 데이터를 인공지능 개발 환경에 맞게 전처리 과정을 거쳐야 한다. 이 연구에서는 크게 데이터 선택, 데이터 결측치 처리, 데이터 인코딩의 과정을 수행하였다.

① 데이터 선택

이 연구의 목적은 한국프로야구 골든글러브 수상자 예측 모델을 개발하는 것이므로, <표 5>는 투수 세이버 메트릭스 기록에 대한 설명이다.

표 5. 투수 세이버 메트릭스 기록 변인 설명

변인	설명	변인	설명
FIP	수비무관 평균자책점	WHIP	이닝당 출루 허용율
BABIP	인플레이 타구 타율	WPA	승리 확률 기여도
LOB%	잔루율	WAR	대체선수 대비 승리기여도
종속변인		GG(골든글러브 수상여부)	

<표 6>은 투수 기본 기록 41개에 대한 내용이다.

표 6. 투수 기본 기록 변인 설명

변인	설명	변인	설명
game	출장 수	BB/9	9이닝 당 사사구 수
IP	이닝 수	K/BB	볼넷 당 탈삼진 수
ERA	평균자책점	HR/9	9이닝 당 피홈런 수
HR	피홈런 수	K%	삼진 비율
BB	볼넷 수	BB%	볼넷 비율
IBB	고의 4구 수	K-BB%	볼넷 하나당 삼진 비율
HBP	사구 수	CYP	사이영상 포인트
K	탈삼진 수	op_bat%	상대 타율
balk	보크 수	op_onbase%	상대 출루율
WP	폭투 수	batter	상대 타자 수
CG	완투 수	hit	피안타 수
SO	완봉 수	2B	2루타 수
start	선발 수	3B	3루타 수
win	승리 수	SLG	상대타자 장타율
P	투구수	OPS	상대타자 OPS
save	세이브 수	lose	패배 수
hold	홀드 수	IP/G	게임당 이닝수
runs	실점 수	P/G	게임당 투구수
ER	자책점 수	K_9	9이닝 당 탈삼진 수
P/IP	이닝당 투구수	P/PA	한 타자당 투구수
PFR	(삼진+볼넷)/이닝		
종속변인		GG(골든글러브 수상여부)	

<표 7>은 타자 세이버 메트릭스 기록에 대한 내용이다.

표 7. 타자 세이버 메트릭스 기록 변인 설명

변인	설명	변인	설명
wOBA	가중 출루율	IsoP	순수 장타율
wRC	조정 득점 창출력	IsoD	순수 출루율
wRC/27	27아웃당 득점 생산력	BABIP	인플레이 타구 타율
WAR	대체선수 대비 승리기여도	wRAA	리그평균대비 득점기여도
종속변인		GG(골든글러브 수상여부)	

<표 8>은 타자 기본기록 29개에 대한 내용이다.

표 8. 타자 기본 기록 변인 설명

변인	설명	변인	설명
game	출장 경기 수	IBB	고의 4구 수
PA	타석 수	SO	삼진 수
AB	타수	GIDP	병살 수
runs	득점 수	SAC	희타 수
HIT	안타 수	SF	희비 수
2B	2루타 수	AVG	타율
3B	3루타 수	OBP	출루율
HR	홈런 수	SLG	장타율
TB	루타 수	OPS	OPS(출루율 + 장타율)
RBI	타점 수	HR%	홈런 비율
SB	도루 수	BB%	사사구 비율
CS	도루 실패 수	K%	삼진 비율
BB	볼넷 수	BB/K	삼진당 볼넷 비율
HBP	사구 수	PSN	호타준족 점수
spd	스피드 스코어		
종속변인		GG(골든글러브 수상여부)	

위의 투수, 타자 기록을 예측 모델 개발 및 적용 데이터로 선정하였다.

② 데이터 결측치 처리

데이터 행렬은 빈칸이 없는 완전행렬로 구성되어야 한다. 따라서 데이터 스크리닝(Data Screening)을 통해 연도별 수집 데이터 차이로 인해 결측치는 Linear regression으로 추정한 후 적용하였다.

③ 데이터 인코딩

인공지능 예측을 환경에 맞도록 정리된 데이터의 인코딩을 변경하였다. 데이터 인코딩은 연속변수는 스케일을 맞춰주는 최대-최소 정규화 과정을 수행하여 데이터를 인코딩하였다. Zscoring값과 min-max값으로 구분하였다.

2) 예측모델 선정

골든글러브 수상자 예측을 위한 로지스틱 회귀분석 및 머신러닝 모델로는 서포트 백터머신(Support vector machine), 랜덤포레스트(Random Forest), XGBoost 모델을 선정하였다. <표 9>는 머신러닝 모델을 나타낸 표이다. 4개의 모델 중 가장 예측이 정확한 모델을 선정하였다.

표 9. 로지스틱 회귀분석 및 머신러닝 모델

구분	모델명
1	로지스틱 회귀분석(Logistic Regression)
2	서포트 백터 머신(Support vector machine)
3	랜덤포레스트(Random Forest)
4	XG부스트(XGBoost)

3) 성능평가

골든글러브 수상자 예측 모델의 성능평가는 정확도(accuracy), 민감도(sensitivity), 특이도(specificity) 정밀도(precision), 재현율(recall), F1 score 지표를 사용하였다.

Ⅳ. 연구결과

이 연구는 2003년부터 2022년까지 한국프로야구 골든글러브 후보 및 수상자의 기록을 기반으로 골든글러브 수상자 예측을 위해 로지스틱 회귀분석(Logistic Regression)과 머신러닝 예측기법 중 서포트 백터머신(Support vector machine), 랜덤포레스트(Random Forest), XGBoost를 대상으로 zscoring값과 minmax값을 활용하여 각 모델별 2가지의 예측모델을 설계하였다. 이후 각 모델의 성능을 평가 및 비교하였다.

1. 골든글러브 후보 및 수상자 간 포지션별 기록 비교

2003년부터 2022년까지 한국프로야구 각 포지션별 골든글러브 후보 및 수상자의 기록을 최소값(Min), 최대값(Max), 평균(Mean), 표준편차(Standard Deviation) 값으로 나타냈다.

1) 투수 변인별 기록

<표 10>과 <표 11>은 투수에 대한 기술통계 분석 결과이다.

표 10. 투수 기본 기록 기술통계

투수	골든글러브 미수상(n=244)				골든글러브 수상(n=20)				전체(n=264)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
game	19	85	37.4	15.9	25	57	30.7	6.5	19	85	36.9	15.5
IP	31	222.2	137.9	47.3	62.2	234.2	179.3	32.1	31	234.2	141.0	47.6
ERA	0.63	6.94	3.41	1.12	1.82	3.55	2.72	0.53	0.63	6.94	3.39	1.01
HR	0	28	11.6	6.6	1	17	10.4	4.2	0	28	11.5	6.5
BB	7	103	43.0	18.4	20	74	46.1	12.9	7	103	43.2	18.1
IBB	0	10	1.2	1.7	0	3	0.8	0.9	0	10	1.2	1.7
HBP	0	25	7.1	5.0	0	25	6.9	5.8	0	25	7.1	5.1
K	14	210	112.9	40.6	52	225	158.2	40.0	14	225	116.3	42.3
balk	0	4	0.4	0.7	0	2	0.4	0.6	0	4	0.4	0.7
WP	0	18	5.1	3.7	2	14	6.0	3.3	0	18	5.2	3.7
CG	0	6	0.4	0.8	0	6	1.9	2.0	0	6	0.5	1.0
SO	0	2	0.2	0.4	0	4	0.8	1.2	0	4	0.2	0.6
start	0	34	20.2	12.3	0	33	26.8	6.6	0	34	20.7	12.1
win	0	20	9.6	4.4	3	22	17.0	4.0	0	22	10.1	4.8
lose	0	15	6.6	3.3	2	8	4.5	1.7	0	15	6.4	3.3
save	0	47	5.7	12.6	0	46	2.5	10.0	0	47	5.4	12.4
hold	0	40	2.9	8.3	0	1	0.1	0.2	0	40	2.7	8.0
runs	4	121	60.5	28.4	17	88	60.6	16.1	4	121	60.5	27.6
ER	4	115	54.8	25.9	16	74	53.6	13.1	4	115	54.7	25.1
batter	135	932	584.4	203.1	249	947	730.3	130.6	135	947	595.4	202.3
hit	27	233	134.2	53.6	50	209	156.1	33.7	27	233	135.9	52.7
K/9	3	13.1	7.5	1.7	5.61	11.66	7.9	1.5	3	13.1	7.6	1.7
BB/9	1.0	5	2.9	0.9	1.3	3.5	2.3	0.6	1.0	5	2.8	0.9
K/BB	0.8	9.1	2.9	1.2	2.0	6.5	3.6	1.1	0.8	9.1	3.0	1.2
HR/9	0	1.7	0.7	0.3	0.1	0.9	0.5	0.2	0	1.7	0.7	0.3
K%	8.2	37.7	20.0	4.9	15.5	31.7	21.8	4.3	8.2	37.7	20.1	4.9
BB%	2.8	12.9	7.5	2.1	3.8	9.6	6.4	1.6	2.8	12.9	7.4	2.1
K-BB%	-1.6	33.6	12.5	5.4	8.8	22.8	15.3	4.3	-1.6	33.6	12.7	5.4
op_bat%	0.137	0.335	0.251	0.032	0.188	0.278	0,234	0.022	0.137	0.335	0.249	0.032
op_onbase%	0.185	0.404	0.315	0.032	0.188	0.278	0.289	0.020	0.185	0.404	0.313	0.032
CYP	60	96.6	74.9	6.1	69.3	85.8	76.6	4.0	60	96.6	75.1	6.0

game: 출장 경기 수, IP: 이닝, ERA: 평균자책점, HR: 홈런, BB: 볼넷, IBB: 고의 4구, HBP : 몸에 맞는볼, K: 탈삼진, balk: 보크, WP: 폭투, CG: 완투, SO: 완봉, start: 선발, win: 승, lose: 패, save: 세, hold: 홀드, runs:실점, ER: 자책점, batter: 상대 타자, hit: 피안타, K/9: 9이닝당 탈삼진 수, BB/9: 9이닝당 볼넷 수, K/BB: 볼넷당 탈삼진 수, HR/9: 9이닝당 홈런 수, K%: 삼진율, BB%: 볼넷 허용율, op_bat%:상대 타율, op_onbase%: 상대 출루율, CYP: 사이영상 포인트

표 11. 투수 세이버 메트릭스 기록 기술통계

투수	골든글러브 미수상(n=244)				골든글러브 수상(n=20)				전체(n=264)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
WAR	-0.71	8.14	3.41	1.71	3.17	9.20	6.41	1.51	-0.71	9.20	3.63	1.91
FIP	0.61	6.25	3.87	0.90	2.15	4.44	3.16	0.56	0.61	6.25	3.81	0.90
BABIP	0.190	0.370	0.300	0.029	0.259	0.344	0.291	0.021	0.190	0.370	0.300	0.029
LOB%	60.0	96.6	74.9	6.1	69.3	85.8	76.6	4.0	60.0	96.6	75.1	6.0
WHIP	0.67	1.79	1.27	0.18	0.95	1.32	1.13	0.11	0.67	1.79	1.26	0.19
WPA	-3.08	4.31	0.68	1.59	1.24	5.65	3.13	1.41	-3.08	5.65	0.80	1.66

WAR: 대체선수 대비 승리기여도, FIP: 수비 무관 평균자책점, BABIP: 인플레이 타구 타율, LOB%: 잔루율, WHIP: 이닝당 출루 허용률, WPA: 승리확률 기여도

2003~2022년 골든글러브 투수 부분 후보 및 수상자를 보면 전체 264명의 선수 중 20명의 골든글러브 수상자 기록과 244명의 후보 선수 기록이 존재한다. 19명의 선발투수와 1명의 구원투수(마무리투수)가 골든글러브를 수상하였다. WAR을 보면 골든글러브 미수상 그룹 최대값 8.14, 평균이 3.41로 확인되었으며, 수상 그룹 최대값 9.20 평균 6.41로 나타났다. ERA의 경우 미수상 그룹 평균 3.41, 수상 그룹 2.72로 나타났으며, 승리 수의 경우 미수상 그룹 9.6승, 수상 그룹 17승으로 확인되었다. 투수의 개인 능력을 살펴보는데 중요한 지표인 삼진과 볼넷, 볼넷 하나당 삼진 수 지표에서는 미수상 그룹 삼진 112.9개, 볼넷 43.0개, K/BB 2.9로 나타났으며, 수상 그룹 삼진 158.2개, 볼넷 46.1개, K/BB 3.6으로 나타났다. 수상그룹보다 미수상 그룹의 평균 볼넷 수가 적은 것은 골든글러브 후보에 구원선수가 포함되었기 때문이며, 9이닝당 볼넷 수를 보면 미수상 그룹 2.9개, 수상 그룹 2.3개로 수상그룹의 볼넷 허용비율이 낮게 나타났다. 또한 이닝당 출루허용율을 나타내는 WHIP 지표와 상대타자 출루율 부분을 보면 미수상 그룹 평균 1.27, 수상 그룹 평균 1.13로 수상 그룹이 이닝당 출루 허용을 적게 했음을 알 수 있다. 상대타자 출루율은 미수상 그룹 평균 0.251, 수상 그룹 평균 0.234로 수상 그룹의 선수들이 상대 타자에게 출루를 허용하는 확률이 낮다고 볼 수 있다. 세이버 메트릭스 지표 중 투수의 고유능력을 평가하는 수비무관 평균자책점 지표인 FIP 지표에서 미수상 그룹 평균 3.87 수상 그룹 평균 3.16로 수상 그룹이 능력이 뛰어남을 알 수 있다.

2) 1루수 변인별 기록

<표 12>는 1루수 기본 기록 기술통계에 대한 결과이다.

표 12. 1루수 기본 기록 기술통계

1루수	골든글러브 미수상(n=79)				골든글러브 수상(n=20)				전체(n=99)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
game	101	143	122.4	10.3	105	142	127.7	9.0	101	143	123.5	10.3
PA	375	622	496.5	58.6	482	627	546.5	43.5	375	627	506.6	59.4
AB	330	536	429.9	50.5	400	540	457.1	35.9	330	540	435.4	49.2
runs	26	129	65.2	18.9	63	130	89.4	19.0	26	130	70.1	21.3
hit	80	181	126.7	21.9	118	180	147.0	17.9	80	181	130.8	22.7
2B	10	42	24.2	7.0	13	42	26.1	7.8	10	42	24.6	7.2
3B	0	13	1.2	2.0	0	5	0.9	1.3	0	13	1.2	1.8
HR	2	53	18.4	9.9	16	56	33.3	10.0	2	53	21.4	11.6
TB	109	377	208.6	50.2	230	373	274.8	33.7	109	377	222.0	54.3
RBI	27	146	77.9	23.5	82	144	106.3	16.4	27	146	83.7	25.0
SB	0	34	5.2	5.4	0	40	6.8	9.2	0	40	5.5	6.4
CS	0	10	2.7	2.4	0	9	2.6	2.5	0	10	2.7	2.4
BB	20	91	53.4	17.4	39	103	74.3	20.1	20	103	57.7	19.8
HBP	0	21	6.7	4.6	0	17	10.1	4.3	0	21	7.4	4.7
IBB	0	10	3.2	2.6	3	25	7.7	4.9	0	25	4.1	3.7
SO	27	161	81.1	28.7	53	142	90.5	24.8	27	161	83.0	28.2
GIDP	3	20	10.3	4.3	2	22	10.8	5.4	2	22	10.4	4.5
SAC	0	15	1.8	2.7	0	1	0.1	0.3	0	15	1.4	2.5
SF	1	11	4.7	2.1	1	9	5.0	2.2	1	11	4.7	2.2
AVG	0.225	0.365	0.294	0.029	0.275	0.381	0.321	0.025	0.225	0.381	0.299	0.030
OBP	0.302	0.474	0.376	0.035	0.349	0.498	0.423	0.031	0.302	0.498	0.386	0.039
SLG	0.327	0.714	0.481	0.079	0.521	0.790	0.603	0.072	0.327	0.790	0.505	0.091
OPS	0.670	1.150	0.857	0.104	0.909	1.288	1.025	0.094	0.670	1.288	0.891	0.123
HR%	0.5	8.5	3.6	1.7	2.6	9.4	6.1	1.8	0.5	9.1	4.1	2.0
BB%	4.6	18.3	10.7	3.0	7.8	17.6	13.5	3.1	4.6	18.3	11.3	3.3
K%	6.5	29.6	16.2	4.9	9.1	26.9	16.7	4.9	6.5	29.6	16.3	4.9
BB/K	0.3	2.2	0.7	0.4	0.3	1.7	0.9	0.3	0.3	2.2	0.8	0.4

game: 출장 수, PA: 타석 수, AB: 타수, runs: 득점, hit: 안타, 2B: 2루타, 3B: 3루타, HR: 홈런, TB: 총루타, RBI: 타점, SB: 도루, CS: 도루 실패, BB: 볼넷, HBP: 몸에 맞는볼, IBB: 고의 4구, SO: 삼진, GIDP: 병살, SAC:희타, SF: 희비, AVG: 타율, OBP: 출루율, SLG: 장타율, OPS: 장타율+출루율, HR%: 홈런률, BB%: 볼넷률, K%: 삼진률, BB/K: 삼진당 볼넷 수

<표 13>은 1루수 세이버 메트릭스 기록 기술통계에 대한 결과이다.

표 13. 1루수 세이버 메트릭스 기록 기술통계

1루수	골든글러브 미수상(n=79)				골든글러브 수상(n=20)				전체(n=99)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
WAR	0.19	7.76	3.03	1.62	3.68	10.71	6.13	1.57	0.19	10.71	3.66	2.03
wOBA	0.308	0.481	0.383	0.038	0.399	0.530	0.446	0.030	0.308	0.530	0.396	0.044
IsoP	0.057	0.371	0.187	0.063	0.174	0.409	0.282	0.067	0.057	0.409	0.206	0.074
IsoD	0.035	0.129	0.082	0.021	0.071	0.130	0.101	0.020	0.035	0.130	0.086	0.022
BABIP	0.235	0.407	0.322	0.033	0.259	0.390	0.335	0.031	0.235	0.407	0.325	0.033
wRC	39.6	156.3	80.5	23.2	84.9	175.7	116.3	19.4	39.6	175.7	87.8	26.7
wRC27	3.5	11.6	6.7	1.8	7.0	15.1	9.6	1.8	3.5	15.1	7.3	2.1
wRAA	-7.5	73.0	18.6	16.2	28.1	96.0	48.6	14.8	-7.5	96.0	24.7	20.0

WAR: 대체선수 대비 승리기여도, wOBA: 가중 출루율, IsoP: 순수장타율, IsoD: 순수 출루율, BABIP: 인플레이 타구 타율, wRC: 조정 득점 창출력, wRC/27: 27아웃당 득점 생산력, wRAA: 리그평균대비 득점 기여도

2003~2022년 골든글러브 1루수 부분 후보 및 수상자를 보면 전체 99명의 선수 중 20명의 골든글러브 수상자 기록과 79명의 후보 선수 기록이 존재한다. 한국프로야구에서 1루수는 전통적으로 거포, 즉 홈런과 장타를 많이 치는 포지션으로 홈런과 루타수를 보면 미수상 그룹 홈런 평균 18.4개, 루타수 208.6로 나타났으며, 수상 그룹 홈런 평균 33.3개, 루타수 274.8로 수상 그룹이 장타력이 뛰어남을 알 수 있다. 대체선수 대비 승리기여도인 WAR은 미수상 그룹 최대값 7.76, 평균 3.03, 수상 그룹 최대값 10.71, 평균 6.13으로 수상 그룹의 WAR이 높게 확인되었다. 세이버 메트릭스 지표 중 순수장타율을 나타내는 IsoP지표와 절대 출루율을 나타내는 IsoD지표는 미수상 그룹 평균 각 0.187, 0.082, 수상 그룹 평균 각 0.282, 0.101로 수상 그룹이 파워와 출루율 모두 높음을 알 수 있다. 조정득점 창출력을 나타내는 wRC 지표는 전체 1루수 포지션에서는 평균 87,8 미수상 그룹 평균 80.5, 수상 그룹 평균 116.3으로 나타났다.

3) 2루수 변인별 기록

<표 14>는 2루수 기본 기록 기술통계에 대한 결과이다.

표 14. 2루수 기본 기록 기술통계

2루수	골든글러브 미수상(n=73)				골든글러브 수상(n=20)				전체(n=93)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
game	104	144	125.8	9.7	105	140	125.8	9.2	104	144	125.8	9.6
PA	339	650	492.7	71.0	415	646	537.6	62.7	339	650	502.4	71.7
AB	298	575	426.9	62.8	355	560	467.8	52.7	298	575	435.7	63.1
runs	31	121	63.6	18.4	54	135	84.8	20.6	31	135	68.2	20.8
hit	72	179	121.5	24.0	104	201	147.0	24.7	72	201	126.9	26.3
2B	7	38	21.4	6.6	18	41	25.8	6.2	7	41	22.3	6.8
3B	0	6	2.2	1.8	0	17	4.3	4.0	0	17	2.6	2.6
HR	0	31	8.7	6.1	1	48	10.5	10.1	0	48	9.1	7.2
TB	90	276	173.4	38.8	135	318	212.6	46.2	90	318	181.8	43.6
RBI	25	98	53.0	16.7	35	137	64.4	25.6	25	137	55.5	19.5
SB	0	46	14.6	11.2	1	53	22.5	15.8	0	53	16.3	12.8
CS	0	20	5.5	3.6	0	17	6.9	4.8	0	20	5.8	3.9
BB	24	96	45.9	14.6	33	105	51.6	18.9	24	105	47.2	15.8
HBP	0	18	6.5	4.4	0	17	7.5	4.2	0	18	6.7	4.4
IBB	0	5	1.1	1.3	0	8	1.7	2.0	0	8	1.2	1.5
SO	28	133	72.9	24.5	39	105	61.1	19.0	28	133	70.4	24.0
GIDF	2	19	9.7	3.7	1	21	9.4	4.7	1	21	9.6	3.9
SAC	0	36	9.5	8.0	0	17	5.8	4.9	0	36	8.7	7.6
SF	0	10	3.9	2.2	1	11	4.9	2.8	0	11	4.1	2.4
AVG	0.231	0.363	0.283	0.028	0.266	0.370	0.313	0.030	0.231	0.370	0.290	0.031
OBP	0.291	0.441	0.359	0.032	0.342	0.438	0.386	0.026	0.291	0.441	0.365	0.032
SLG	0.301	0.552	0.404	0.057	0.367	0.596	0.451	0.063	0.301	0.596	0.414	0.062
OPS	0.600	0.969	0.762	0.081	0.709	0.988	0.837	0.079	0.600	0.988	0.778	0.086
HR%	0.0	5.2	1.7	1.2	0.2	7.5	1.9	1.6	0.0	7.5	1.8	1.3
BB%	5.6	16.0	9.3	2.2	6.6	17.3	9.5	2.7	5.6	17.3	9.4	2.3
K%	6.2	26.3	14.8	5.0	7.6	19.6	11.4	3.2	6.2	26.3	14.0	4.9
BB/K	0.3	1.4	0.7	0.3	0.5	1.3	0.9	0.3	0.3	1.4	0.7	0.3

game: 출장 수, PA: 타석 수, AB: 타수, runs: 득점, hit: 안타, 2B: 2루타, 3B: 3루타, HR: 홈런, TB: 총 루타, RBI: 타점, SB: 도루, CS: 도루 실패, BB: 볼넷, HBP: 몸에 맞는볼, IBB: 고의 4구, SO: 삼진, GIDP: 병살, SAC:희타, SF: 희비, AVG: 타율, OBP: 출루율, SLG: 장타율, OPS: 장타율+출루율, HR%: 홈런률, BB%: 볼넷률, K%: 삼진률, BB/K: 삼진당 볼넷 수

<표 15>는 2루수 세이버 메트릭스 기록 기술통계에 대한 결과이다.

표 15. 2루수 세이버 메트릭스 기록 기술통계

2루수	골든글러브 미수상(n=73)				골든글러브 수상(n=20)				전체(n=93)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
WAR	0.03	6.23	2.66	1.30	2.80	7.63	4.66	1.24	0.03	7.63	3.09	1.53
wOBA	0.269	0.424	0.344	0.034	0.328	0.437	0.378	0.028	0.269	0.437	0.351	0.035
IsoP	0.053	0.245	0.120	0.044	0.085	0.309	0.138	0.052	0.053	0.309	0.124	0.046
IsoD	0.046	0.121	0.076	0.016	0.050	0.124	0.073	0.020	0.046	0.124	0.075	0.017
BABIP	0.259	0.408	0.322	0.030	0.248	0.395	0.339	0.035	0.248	0.408	0.326	0.032
wRC	27.8	121.0	64.2	19.6	55.7	131.4	85.0	21.9	27.8	131.4	68.6	21.8
wRC27	2.3	9.0	5.2	1.5	4.3	9.7	6.6	1.4	2.3	9.7	5.5	1.6
wRAA	-21.2	36.2	2.2	11.9	-2.8	44.6	18.2	11.9	-21.2	44.6	5.7	13.6

WAR: 대체선수 대비 승리기여도, wOBA: 가중 출루율, IsoP: 순수장타율, IsoD: 순수 출루율, BABIP: 인플레이 타구 타율, wRC: 조정 득점 창출력, wRC/27: 27아웃당 득점 생산력, wRAA: 리그평균대비 득점 기여도

2003-2022 골든글러브 2루수 부분 후보 및 수상자를 보면 전체 93명의 선수 중 20명의 골든글러브 수상자 기록과 73명의 후보 선수 기록이 존재한다. 2루수의 기록에서 타율을 보면 전체 평균 0.290, 미수상 그룹 평균 0.283, 수상 그룹 평균 0.313으로 나타났다. 희생타와 희비 수를 보면 미수상 그룹이 평균 9.5개, 3.9개로 나타났으며, 수상 그룹 5.8개, 4.9개로 확인되었는데, 희생 번트의 수는 미수상 그룹이 높게 나타났지만 희생플라이 즉 희비의 수는 수상그룹이 높게 나타났다. 대체선수 대비 승리기여도인 WAR은 미수상 그룹 평균 2.66, 수상자 그룹 4.66으로 타 포지션에 비해 WAR지표의 수치가 낮게 나타났음을 알 수 있다.

4) 3루수 변인별 기록

<표 16>은 3루수 기본 기록 기술통계에 대한 결과이다.

표 16. 3루수 기본 기록 기술통계

3루수	골든글러브 미수상(n=78)				골든글러브 수상(n=20)				전체(n=98)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
game	99	144	122.1	9.6	109	141	126.7	9.1	99	144	123.0	9.7
PA	380	601	488.2	47.6	425	606	530.3	48.2	380	606	496.8	50.7
AB	332	547	426.0	44.4	356	541	448.4	49.3	332	547	430.5	46.3
runs	35	97	61.6	14.2	53	108	79.3	15.3	35	108	65.2	16.1
hit	80	167	122.4	17.5	108	174	135.5	18.9	80	174	125.0	18.6
2B	7	41	22.0	5.4	12	35	23.9	6.1	7	41	22.4	5.6
3B	0	5	1.3	1.3	0	5	1.2	1.4	0	5	1.3	1.4
HR	1	35	13.7	7.6	10	46	26.5	9.4	1	46	16.3	9.5
TB	126	284	188.1	34.0	182	319	241.2	36.1	126	319	198.9	40.6
RBI	34	113	66.6	17.5	68	133	92.4	18.9	34	133	71.8	20.6
SB	0	39	6.4	7.2	0	24	7.5	7.3	0	39	6.6	7.2
CS	0	11	3.1	2.8	0	8	3.6	2.2	0	11	3.2	2.7
BB	18	75	45.7	13.9	29	84	58.3	16.9	18	84	48.2	15.4
HBP	0	27	8.1	5.8	5	26	16.1	6.2	0	27	9.7	6.7
IBB	0	10	2.0	2.1	0	11	3.6	3.4	0	11	2.3	2.5
SO	28	129	71.8	19.3	52	126	85.9	20.7	28	129	74.7	20.4
GIDF	3	21	11.7	3.8	4	23	11.0	4.5	3	23	11.5	4.0
SAC	0	16	4.1	4.1	0	13	2.0	3.1	0	16	3.6	4.0
SF	0	12	4.4	2.5	1	13	5.7	3.1	0	13	4.6	2.7
AVG	0.222	0.353	0.287	0.026	0.257	0.364	0.302	0.025	0.222	0.364	0.290	0.027
OBP	0.296	0.455	0.363	0.031	0.335	0.457	0.398	0.032	0.296	0.457	0.370	0.034
SLG	0.317	0.593	0.441	0.062	0.432	0.684	0.539	0.066	0.317	0.684	0.461	0.074
OPS	0.618	1.049	0.805	0.087	0.767	1.111	0.937	0.090	0.618	1.111	0.832	0.102
HR%	0.2	7.2	2.8	1.5	1.7	8.7	5.0	1.7	0.2	8.7	3.2	1.8
BB%	3.9	16.7	9.4	2.7	5.4	17.3	11.1	3.3	3.9	17.3	9.7	2.9
K%	5.8	26.4	14.8	3.8	9.0	20.8	16.2	3.5	5.8	26.4	15.0	3.8
BB/K	0.2	1.3	0.7	0.3	0.4	1.5	0.7	0.3	0.2	1.5	0.7	0.3

game: 출장 수, PA: 타석 수, AB: 타수, runs: 득점, hit: 안타, 2B: 2루타, 3B: 3루타, HR: 홈런, TB: 총루타, RBI: 타점, SB: 도루, CS: 도루 실패, BB: 볼넷, HBP: 몸에 맞는볼, IBB: 고의 4구, SO: 삼진, GIDP: 병살, SAC:희타, SF: 희비, AVG: 타율, OBP: 출루율, SLG: 장타율, OPS: 장타율+출루율, HR%: 홈런률, BB%: 볼넷률, K%: 삼진률, BB/K: 삼진당 볼넷 수

<표 17>은 3루수 세이버 메트릭스 기록 기술통계에 대한 결과이다.

표 17. 3루수 세이버 메트릭스 기록 기술통계

3루수	골든글러브 미수상(n=78)				골든글러브 수상(n=20)				전체(n=98)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
WAR	0.38	7.89	3.28	1.42	3.04	8.82	5.63	1.37	0.38	8.82	3.76	1.70
wOBA	0.289	0.454	0.363	0.034	0.346	0.482	0.414	0.034	0.289	0.482	0.373	0.040
IsoP	0.058	0.303	0.154	0.052	0.136	0.367	0.237	0.056	0.058	0.367	0.171	0.063
IsoD	0.029	0.130	0.076	0.023	0.052	0.135	0.095	0.025	0.029	0.135	0.080	0.025
BABIP	0.250	0.379	0.315	0.031	0.277	0.365	0.318	0.027	0.250	0.379	0.316	0.030
wRC	38.7	107.7	70.1	16.5	70.2	135.9	98.1	17.4	38.7	135.9	75.8	20.1
wRC27	3.1	10.3	5.8	1.4	4.8	11.3	8.0	1.6	3.1	11.3	6.3	1.7
wRAA	-18.9	47.3	9.7	13.4	2.9	65.6	32.4	14.5	-18.9	65.6	14.3	16.4

WAR: 대체선수 대비 승리기여도, wOBA: 가중 출루율, IsoP: 순수장타율, IsoD: 순수 출루율, BABIP: 인플레이 타구 타율, wRC: 조정 득점 창출력, wRC/27: 27아웃당 득점 생산력, wRAA: 리그평균대비 득점 기여도

2003~2022년 골든글러브 3루수 부분 후보 및 수상자를 보면 전체 98명의 선수 중 20명의 골든글러브 수상자 기록과 78명의 후보 선수 기록이 존재한다. 한국프로야구에서 3루수는 1루수처럼 장타력이 있는 선수들이 많은 포지션이다. WAR을 보면 전체 평균 3.76, 미수상 그룹 평균 3.28, 수상 그룹 평균 5.63으로 확인되었다. 홈런과 장타율은 미수상 그룹 평균 13.7개, 0.441, 수상 그룹 평균 26.5개, 0.539로 수상 그룹의 홈런 수와 장타율 모두 높게 나타났다. K%와 BB%를 보면 미수상 그룹 14.8%, 9.4%, 수상 그룹 16.2%, 11.1%로 확인되었는데, 이는 수상 그룹이 삼진도 많이 당하고 볼넷도 많이 나가고 있음을 알 수 있다. 출루율+장타율 지표인 OPS는 미수상 그룹 0.805, 수상 그룹 0.937로 수상 그룹이 0.132가 높게 나타났다.

5) 유격수 변인별 기록

<표 18>은 유격수 기본 기록 기술통계에 대한 결과이다.

표 18. 유격수 기본 기록 기술 통계

유격수	골든글러브 미수상(n=78)				골든글러브 수상(n=20)				전체(n=98)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
game	98	144	126.2	10.9	100	144	127.7	10.7	98	144	126.5	10.9
PA	167	617	472.7	71.6	400	634	515.0	66.6	167	634	481.4	72.6
AB	149	533	414.1	63.8	333	559	443.7	61.6	149	559	420.1	64.5
runs	22	95	58.4	17.8	35	112	72.6	21.5	22	112	61.3	19.5
hit	37	159	112.8	21.4	102	176	134.3	21.6	37	176	117.1	23.1
2B	5	41	19.8	7.6	13	38	24.5	6.8	5	41	20.7	7.6
3B	0	8	2.3	2.1	0	5	2.0	1.4	0	8	2.3	2.0
HR	0	23	7.7	5.8	1	40	14.6	10.3	0	40	9.1	7.5
TB	48	270	160.2	41.2	120	309	206.5	47.3	48	309	169.7	46.4
RBI	8	114	49.5	16.4	36	117	74.5	21.6	8	117	54.6	20.3
SB	0	53	13.3	10.2	2	46	12.5	11.1	0	53	13.2	10.4
CS	0	15	5.1	3.2	2	8	4.5	2.0	0	15	5.0	3.0
BB	6	67	39.6	13.4	31	75	55.3	11.9	6	75	42.8	14.5
HBP	0	14	5.7	2.9	1	13	6.0	2.9	0	14	5.7	2.9
IBB	0	4	0.6	0.9	0	8	2.2	2.1	0	8	0.9	1.4
SO	26	146	72.9	27.6	38	109	70.9	21.8	26	146	72.4	26.5
GIDF	1	20	8.5	3.9	5	18	10.4	4.2	1	20	8.9	4.0
SAC	1	25	9.3	5.7	0	14	4.8	4.0	0	25	8.4	5.7
SF	0	11	4.0	2.2	2	13	5.4	2.7	0	13	4.3	2.4
AVG	0.221	0.350	0.271	0.023	0.269	0.370	0.303	0.024	0.221	0.370	0.277	0.026
OBP	0.274	0.395	0.340	0.028	0.347	0.459	0.383	0.027	0.274	0.459	0.349	0.033
SLG	0.273	0.536	0.383	0.061	0.342	0.739	0.464	0.081	0.273	0.739	0.400	0.073
OPS	0.548	0.908	0.723	0.081	0.706	1.198	0.847	0.102	0.548	1.198	0.748	0.099
HR%	0.0	4.6	1.6	1.2	0.3	8.0	2.8	1.9	0.0	8.0	1.8	1.4
BB%	2.7	14.3	8.4	2.3	7.4	13.7	10.8	1.9	2.7	14.3	8.8	2.4
K%	5.8	32.9	15.4	5.2	7.6	21.2	13.7	3.7	5.8	32.9	15.1	5.0
BB/K	0.1	1.7	0.6	0.3	0.5	1.4	0.8	0.2	0.1	1.7	0.7	0.3

game: 출장 수, PA: 타석 수, AB: 타수, runs: 득점, hit: 안타, 2B: 2루타, 3B: 3루타, HR: 홈런, TB: 총루타, RBI: 타점, SB: 도루, CS: 도루 실패, BB: 볼넷, HBP: 몸에 맞는볼, IBB: 고의 4구, SO: 삼진, GIDP: 병살, SAC:희타, SF: 희비, AVG: 타율, OBP: 출루율, SLG: 장타율, OPS: 장타율+출루율, HR%: 홈런률, BB%: 볼넷률, K%: 삼진률, BB/K: 삼진당 볼넷수

<표 19>는 유격수 세이버 메트릭스 기록 기술통계에 대한 결과이다.

표 19. 유격수 세이버 메릭스 기록 기술통계

유격수	골든글러브 미수상(n=78)				골든글러브 수상(n=20)				전체(n=98)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
WAR	-1.13	5.61	2.34	1.37	2.52	8.23	4.81	1.64	-1.13	8.23	2.84	1.74
wOBA	0.254	0.401	0.327	0.032	0.334	0.500	0.381	0.037	0.254	0.500	0.338	0.039
IsoP	0.032	0.229	0.112	0.048	0.051	0.383	0.161	0.071	0.032	0.383	0.122	0.057
IsoD	0.028	0.111	0.069	0.017	0.050	0.103	0.081	0.014	0.028	0.111	0.071	0.017
BABIP	0.253	0.384	0.311	0.026	0.283	0.398	0.330	0.028	0.253	0.398	0.315	0.028
wRC	12.3	101.2	54.6	17.7	46.4	136.1	81.7	21.3	12.3	136.1	60.1	21.5
wRC27	1.9	7.1	4.5	1.2	4.6	13.0	6.6	1.8	1.9	13.0	4.9	1.6
wRAA	-40.8	27.3	-4.2	11.9	-0.4	65.5	17.8	16.3	-40.8	65.5	0.3	15.7

WAR: 대체선수 대비 승리기여도, wOBA: 가중 출루율, IsoP: 순수장타율, IsoD: 순수 출루율, BABIP: 인플레이 타구 타율, wRC: 조정 득점 창출력, wRC/27: 27아웃당 득점 생산력, wRAA: 리그평균대비 득점기여도

2003~2022년 골든글러브 유격수 부분 후보 및 수상자를 보면 전체 98명의 선수 중 20명의 골든글러브 수상자 기록과 78명의 후보 선수 기록이 존재한다. 유격수 포지션은 타 포지션에 비해 공격보다 수비가 더 중요한 포지션으로 인식되고 있다. WAR을 보면 전체 평균 2.84, 미수상 그룹 평균 2.34, 수상 그룹 평균 4.81로 수상 그룹이 높게 나타났음을 알 수 있다. 타율은 미수상 그룹 평균 0.271, 수상 그룹 0.303으로 나타났으며, 출루율은 미수상 그룹 평균 0.340, 수상 그룹 평균 0.383로 수상 그룹이 높은 수치를 보였다. 조정득점 창출력을 나타내는 wRC 지표는 전체 평균 60.1, 미수상 그룹 평균 54.6, 수상 그룹 평균 81.7으로 확인되었다. 수비가 중요한 유격수 포지션에서 타격지표가 높게 나타난 이유 중 하나로 메이저리그에 진출한 유격수 강정호, 김하성 선수의 기록이 포함되어 전체적인 기록이 높게 나타났음을 알 수 있다.

6) 포수 변인별 기록

<표 20>은 포수 기본 기록 기술통계에 대한 결과이다.

표 20. 포수 기본 기록 기술통계

포수	골든글러브 미수상(n=69)				골든글러브 수상(n=20)				전체(n=89)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
game	99	139	122.0	10.2	97	133	120.9	10.2	97	139	121.7	10.2
PA	204	526	413.7	60.7	338	552	464.1	54.2	204	552	425.0	62.9
AB	181	451	361.3	53.8	296	501	401.1	53.5	181	501	370.2	56.3
runs	22	94	42.7	11.5	28	86	56.2	15.3	22	94	45.7	13.6
hit	49	134	97.5	18.3	77	165	119.8	24.8	49	165	102.5	22.0
2B	4	33	16.2	5.2	13	29	22.0	4.5	4	33	17.5	5.6
3B	0	9	0.8	1.4	0	2	0.6	0.7	0	9	0.7	1.3
HR	0	35	10.3	7.4	6	33	17.7	6.6	0	35	12.0	7.8
TB	70	244	146.1	38.2	123	278	195.9	44.4	70	278	157.3	44.8
RBI	18	86	50.8	15.5	41	124	72.4	20.0	18	124	55.7	18.9
SB	0	8	1.9	2.0	0	10	3.2	2.4	0	10	2.2	2.1
CS	0	10	1.7	1.9	0	7	2.3	1.9	0	10	1.8	1.9
BB	12	94	34.7	15.1	24	60	43.8	8.7	12	94	36.8	14.4
HBP	1	27	7.7	5.4	3	24	10.1	4.8	1	27	8.2	5.3
IBB	0	7	1.2	1.3	0	8	3.4	2.2	0	8	1.7	1.8
SO	28	126	63.1	20.2	29	104	61.9	21.1	28	126	62.8	20.4
GIDF	4	21	10.9	4.5	5	20	12.4	4.4	4	21	11.3	4.5
SAC	0	21	6.9	4.8	0	15	3.6	4.6	0	21	6.1	4.9
SF	0	7	3.1	1.7	1	10	5.6	2.4	0	10	3.7	2.1
AVG	0.222	0.329	0.269	0.022	0.230	0.358	0.297	0.034	0.222	0.358	0.275	0.028
OBP	0.278	0.440	0.342	0.030	0.316	0.438	0.376	0.028	0.278	0.440	0.350	0.033
SLG	0.261	0.639	0.400	0.068	0.361	0.603	0.485	0.068	0.261	0.639	0.419	0.076
OPS	0.539	1.061	0.742	0.090	0.678	1.013	0.861	0.092	0.539	1.061	0.769	0.103
HR%	0.0	7.7	2.4	1.6	1.5	6.3	3.7	1.2	0.0	7.7	2.7	1.6
BB%	2.9	17.9	8.3	3.1	6.5	14.8	9.5	1.9	2.9	17.9	8.6	3.0
K%	7.0	24.6	15.0	4.0	7.4	21.5	13.3	4.3	7.0	24.6	14.6	4.1
BB/K	0.2	1.3	0.6	0.2	0.4	1.4	0.8	0.3	0.2	1.4	0.6	0.2

game: 출장 수, PA: 타석 수, AB: 타수, runs: 득점, hit: 안타, 2B: 2루타, 3B: 3루타, HR: 홈런, TB: 총루타, RBI: 타점, SB: 도루, CS: 도루 실패, BB: 볼넷, HBP: 몸에 맞는볼, IBB: 고의 4구, SO: 삼진, GIDP: 병살, SAC:희타, SF: 희비, AVG: 타율, OBP: 출루율, SLG: 장타율, OPS: 장타율+출루율, HR%: 홈런률, BB%: 볼넷률, K%: 삼진률, BB/K: 삼진당 볼넷수

<표 21>은 포수 세이버 메트릭스 기록 기술통계에 대한 결과이다.

표 21. 포수 세이버 메트릭스 기록 기술통계

포수	골든글러브 미수상(n=69)				골든글러브 수상(n=20)				전체(n=89)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
WAR	-0.54	8.36	2.44	1.44	1.24	6.61	4.54	1.42	-0.54	8.36	2.91	1.68
wOBA	0.253	0.451	0.335	0.036	0.304	0.452	0.383	0.035	0.253	0.452	0.346	0.041
IsoP	0.022	0.329	0.130	0.057	0.101	0.276	0.188	0.044	0.022	0.329	0.143	0.060
IsoD	0.029	0.145	0.073	0.025	0.055	0.131	0.079	0.018	0.029	0.145	0.075	0.024
BABIP	0.246	0.355	0.300	0.025	0.251	0.360	0.310	0.029	0.246	0.360	0.302	0.026
wRC	13.9	111.9	50.5	17.5	39.8	109.6	74.8	20.1	13.9	111.9	56.0	20.7
wRC27	1.7	10.0	4.7	1.5	3.4	9.8	6.6	1.6	1.7	10.0	5.2	1.7
wRAA	-30	48.5	-0.79	13.3	-18.1	44.3	17.4	14.5	-30	48.5	3.3	15.5

WAR: 대체선수 대비 승리기여도, wOBA: 가중 출루율, IsoP: 순수장타율, IsoD: 순수 출루율, BABIP: 인플레이 타구 타율, wRC: 조정 득점 창출력, wRC/27: 27아웃당 득점 생산력, wRAA: 리그평균대비 득점 기여도

2003~2022년 골든글러브 포수 부분 후보 및 수상자를 보면 전체 89명의 선수 중 20명의 골든글러브 수상자 기록과 69명의 후보 선수 기록이 존재한다. WAR은 전체 평균 2.91, 미수상 그룹 평균 2.44, 수상 그룹 평균 4.54로 나타났다. 체력적 부담이 큰 포수 포지션의 경기 출장 수를 보면 미수상 그룹 평균 122.0경기, 수상 그룹 평균 120.9경기로 20경기 내외를 결장했음을 알 수 있다. 타율은 미수상 그룹 평균 0.269, 수상 그룹 0.297로 타 포지션에 타율은 낮음을 확인할 수 있었다. 병살 수는 미수상 그룹 평균 10.9개, 수상 그룹 평균 12.4개로 나타났는데 포수 포지션의 선수가 스피드가 타 포지션에 비해 떨어지기 때문에 병살 수가 많음을 알 수 있다.

7) 외야수 변인별 기록

<표 22>는 외야수 기본 기록 기술통계에 대한 결과이다.

표 22. 외야수 기본 기록 기술통계

외야수	골든글러브 미수상(n=235)				골든글러브 수상(n=61)				전체(n=296)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
game	85	144	125.4	12.1	109	144	129.7	10.0	85	144	126.3	11.8
PA	333	672	514.5	76.1	450	667	566.3	50.4	333	672	525.2	74.6
AB	285	600	450.1	68.7	401	576	490.9	45.5	285	600	458.5	66.7
runs	28	118	72.6	18.2	58	118	89.3	14.9	28	118	76.0	18.9
hit	60	192	133.7	26.3	110	195	160.9	19.8	60	195	139.3	27.4
2B	5	47	22.7	7.7	13	49	28.8	8.4	5	49	24.0	8.2
3B	0	12	2.7	2.5	0	12	3.1	2.8	0	12	2.8	2.5
HR	0	43	11.0	8.8	0	53	19.7	12.2	0	53	12.8	10.2
TB	77	333	194.9	48.8	157	374	254.9	51.5	77	374	207.3	55.0
RBI	13	123	60.4	23.2	28	144	86.8	29.4	13	144	65.9	26.8
SB	0	66	15.0	12.9	0	53	15.2	14.2	0	66	15.0	13.2
CS	0	21	5.9	4.2	0	20	5.5	3.8	0	21	5.8	4.1
BB	13	92	49.2	15.1	30	124	61.6	20.6	13	124	51.7	17.1
HBP	0	20	6.3	4.4	0	17	6.6	3.7	0	20	6.4	4.2
IBB	0	10	1.9	2.2	0	17	4.2	3.8	0	17	2.4	2.8
SO	29	156	75.5	25.2	32	137	74.8	26.8	29	156	75.3	25.6
GIDF	1	23	8.5	4.0	3	19	9.1	3.7	1	23	8.6	3.9
SAC	0	24	4.9	5.1	0	13	2.2	3.2	0	24	4.3	4.9
SF	0	12	4.0	2.4	1	12	5.0	2.7	0	12	4.2	2.5
AVG	0.189	0.366	0.296	0.026	0.258	0.376	0.328	0.023	0.189	0.376	0.302	0.029
OBP	0.298	0.438	0.370	0.027	0.353	0.478	0.406	0.029	0.298	0.478	0.378	0.031
SLG	0.243	0.603	0.430	0.069	0.332	0.720	0.518	0.084	0.243	0.720	0.448	0.080
OPS	0.580	1.032	0.800	0.085	0.709	1.197	0.923	0.101	0.580	1.197	0.826	0.102
HR%	0.0	7.3	2.1	1.6	0.0	8.8	3.4	2.1	0.0	8.8	2.4	1.8
BB%	3.1	16.9	9.5	2.5	5.1	20.6	10.8	3.3	3.1	20.6	9.8	2.8
K%	5.5	27.6	14.7	4.4	5.1	23.9	13.2	4.4	5.1	27.6	14.4	4.4
BB/K	0.2	2.2	0.7	0.3	0.3	2.1	0.9	0.4	0.2	2.2	0.8	0.3

game: 출장 수, PA: 타석 수, AB: 타수, runs: 득점, hit: 안타, 2B: 2루타, 3B: 3루타, HR: 홈런, TB: 총루타, RBI: 타점, SB: 도루, CS: 도루 실패, BB: 볼넷, HBP: 몸에 맞는볼, IBB: 고의 4구, SO: 삼진, GIDP: 병살, SAC:희타, SF: 희비, AVG: 타율, OBP: 출루율, SLG: 장타율, OPS: 장타율+출루율, HR%: 홈런률, BB%: 볼넷률, K%: 삼진률, BB/K: 삼진당 볼넷수

<표 23>은 외야수 세이버 메트릭스 기록 기술통계에 대한 결과이다.

표 23. 외야수 세이버 메트릭스 기록 기술통계

외야수	골든글러브 미수상(n=235)				골든글러브 수상(n=61)				전체(n=296)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
WAR	-0.79	7.46	3.11	1.39	3.35	10.19	5.78	1.32	-0.79	10.19	3.66	1.75
wOBA	0.287	0.440	0.363	0.031	0.334	0.498	0.411	0.034	0.287	0.498	0.373	0.037
IsoP	0.017	0.317	0.135	0.057	0.040	0.385	0.190	0.077	0.017	0.385	0.146	0.066
IsoD	0.033	0.147	0.074	0.020	0.040	0.143	0.078	0.022	0.033	0.147	0.075	0.020
BABIP	0.238	0.413	0.333	0.029	0.277	0.404	0.351	0.029	0.238	0.413	0.337	0.030
wRC	29.0	138.5	74.9	20.6	57.5	154.7	104.7	22.8	29.0	154.7	81.1	24.3
wRC27	2.9	10.1	5.9	1.3	4.4	12.5	8.1	1.7	2.9	12.5	6.4	1.7
wRAA	-26.2	52.6	10.4	13.3	1.3	80.7	34.6	16.2	-26.2	80.7	15.4	17.1

WAR: 대체선수 대비 승리기여도, wOBA: 가중 출루율, IsoP: 순수장타율, IsoD: 순수 출루율, BABIP: 인플레이 타구 타율, wRC: 조정 득점 창출력, wRC/27: 27아웃당 득점 생산력, wRAA: 리그평균대비 득점 기여도

2003-2022 골든글러브 외야수 부분 후보 및 수상자를 보면 전체 296명의 선수 중 61명의 골든글러브 수상자 기록과 235명의 후보 선수 기록이 존재한다(2004년 골든글러브 외야수 부분 4명 수상). WAR을 보면 전체선수 평균 3.66, 미수상 그룹 평균 3.11, 수상 그룹 평균 5.78로 나타났다. 타율과 OPS를 보면 미수상 그룹 평균 0.296, 0.800, 수상 그룹 평균 0.328, 0.923으로 수상 그룹 평균이 타율 0.032, OPS 0.123이 높게 나타났음을 확인할 수 있다. 도루 수를 보면 미수상 그룹 평균 15.0개 수상 그룹 평균 15.2개로 나타났는데 수비 범위가 넓은 외야 포지션에 발이 빠른 선수가 많아 타 포지션에 비해 도루가 높게 나타났음을 알 수 있다.

8) 지명타자 변인별 기록

<표 24>는 지명타자 기본 기록 기술통계에 대한 결과이다.

표 24. 지명타자 기본 기록 기술통계

DH	골든글러브 미수상(n=65)				골든글러브 수상(n=20)				전체(n=85)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
game	93	144	123.4	12.4	98	144	127.8	12.5	93	144	124.4	12.6
PA	379	668	502.1	69.1	401	652	540.5	62.3	379	668	511.2	69.5
AB	317	586	432.4	62.9	374	572	467.3	56.1	317	586	440.6	63.1
runs	38	115	62.4	17.8	39	94	73.2	15.5	38	115	64.9	17.9
hit	86	199	128.2	23.5	125	197	155.9	21.8	86	199	134.7	26.0
2B	11	39	21.9	6.1	17	39	27.7	5.9	11	39	23.3	6.5
3B	0	6	0.8	1.3	0	3	1.0	1.0	0	6	0.8	1.2
HR	4	38	18.5	7.3	5	37	19.7	8.8	4	38	18.7	7.7
TB	140	313	207.1	40.6	170	322	244.4	43.8	140	322	215.9	44.3
RBI	45	123	78.1	18.6	63	136	90.6	20.8	45	136	81.0	19.9
SB	0	25	3.6	5.3	0	20	4.2	4.8	0	25	3.7	5.2
CS	0	10	1.9	2.2	0	8	2.2	2.2	0	10	2.0	2.2
BB	19	108	58.0	18.8	22	108	60.9	22.9	19	108	58.7	19.9
HBP	0	23	6.5	4.5	0	16	6.4	4.2	0	23	6.5	4.4
IBB	0	12	3.2	2.7	0	15	5.7	3.6	0	15	3.8	3.1
SO	35	148	77.6	22.3	35	101	64.7	20.0	35	148	74.5	22.5
GIDP	5	34	12.8	5.7	9	26	13.7	4.5	5	34	13.0	5.5
SAC	0	6	0.8	1.5	0	4	0.6	1.1	0	6	0.7	1.4
SF	1	11	4.4	2.5	1	13	5.4	2.9	1	13	4.6	2.7
AVG	0.233	0.346	0.296	0.022	0.303	0.371	0.333	0.019	0.233	0.371	0.304	0.027
OBP	0.306	0.468	0.384	0.029	0.358	0.476	0.413	0.032	0.306	0.476	0.391	0.032
SLG	0.375	0.596	0.478	0.054	0.403	0.601	0.521	0.057	0.375	0.601	0.488	0.058
OPS	0.730	1.043	0.862	0.072	0.779	1.045	0.934	0.076	0.730	1.045	0.879	0.079
HR%	0.8	6.9	3.7	1.3	1.1	6.1	3.6	1.5	0.8	6.9	3.6	1.4
BB%	5.0	19.5	11.6	3.4	5.4	19.3	11.2	3.9	5.0	19.5	11.5	3.5
K%	5.7	25.3	15.6	4.1	7.6	18.2	11.9	3.0	5.7	25.3	14.7	4.2
BB/K	0.3	1.9	0.8	0.3	0.5	2.4	1.0	0.5	0.3	2.4	0.8	0.4

game: 출장 수, PA: 타석 수, AB: 타수, runs: 득점, hit: 안타, 2B: 2루타, 3B: 3루타, HR: 홈런, TB: 총루타, RBI: 타점, SB: 도루, CS: 도루 실패, BB: 볼넷, HBP: 몸에 맞는볼, IBB: 고의 4구, SO: 삼진, GIDP: 병살, SAC:희타, SF: 희비, AVG: 타율, OBP: 출루율, SLG: 장타율, OPS: 장타율+출루율, HR%: 홈런률, BB%: 볼넷률, K%: 삼진률, BB/K: 삼진당 볼넷수

<표 25>는 지명타자 세이버 메트릭스 기록 기술통계에 대한 결과이다.

표 25. 지명타자 세이버 메트릭스 기록 기술통계

DH	골든글러브 미수상(n=65)				골든글러브 수상(n=20)				전체(n=85)			
	Min	Max	mean	SD	Min	Max	mean	SD	Min	Max	mean	SD
WAR	0.2	5.56	2.85	1.23	2.13	6.72	4.31	1.35	0.2	6.72	3.19	1.40
wOBA	0.339	0.454	0.388	0.026	0.368	0.463	0.417	0.028	0.339	0.463	0.395	0.029
IsoP	0.082	0.303	0.183	0.049	0.097	0.260	0.188	0.053	0.082	0.303	0.184	0.050
IsoD	0.044	0.144	0.088	0.026	0.036	0.143	0.079	0.029	0.036	0.144	0.086	0.027
BABIP	0.253	0.395	0.322	0.031	0.308	0.410	0.351	0.029	0.253	0.410	0.329	0.033
wRC	54.3	123.6	82.8	18.1	63.2	152.8	102.2	21.2	54.3	152.8	87.4	20.6
wRC27	4.5	11.0	6.9	1.3	5.7	11.7	8.3	1.5	4.5	11.7	7.2	1.5
wRAA	-0.6	45.7	20.3	11.4	13.4	61.4	35.2	13.9	-0.6	61.4	23.8	13.6

WAR: 대체선수 대비 승리기여도, wOBA: 가중 출루율, IsoP: 순수장타율, IsoD: 순수 출루율, BABIP: 인플레이 타구 타율, wRC: 조정 득점 창출력, wRC/27: 27아웃당 득점 생산력, wRAA: 리그평균대비 득점 기여도

2003-2022 골든글러브 지명타자 부분 후보 및 수상자를 보면 전체 85명의 선수 중 20명의 골든글러브 수상자 기록과 65명의 후보 선수 기록이 존재한다. 지명타자는 수비를 참여하지 않는 포지션으로 최근 야구에서는 지명타자 자리를 수비를 한 번씩 쉬면서 체력안배를 하는 포지션으로 활용되고 있어 후보 수가 점점 줄어드는 포지션이다. 타격능력을 극대화하는 포지션으로 전문 지명타자들이 후보로 선정되곤 하는데 경기 수를 보면 전체 평균 124.4경기 미수상 그룹 평균 123.4경기 수상 그룹 평균 127.8경기를 출장하고 있다. 타율과 OPS를 보면 미수상 그룹 평균 0.296, 0.862, 수상 그룹 평균 0.333, 0.934로 나타났다. WAR을 보면 미수상 그룹 평균 2.85, 수상 그룹 평균 4.31로 확인되었는데 수비에 대한 점수가 반영되지 않아 전체적으로 WAR 수치가 낮게 나타났음을 알 수 있다.

2. 로지스틱 회귀분석 및 머신러닝 예측모델 분석 결과

1) 로지스틱 회귀분석 및 머신러닝 예측 모델별 최적 변수 결정

(1) 로지스틱 회귀분석

표 26. 로지스틱 회귀분석 파라미터 설명

parameter	value
penalty	L1, L2, elasticnet
L1_ratio	0.1~0.9
max_iter	3000
solver	L1:liblinear L2:lbfgs elasticnet:saga
kind	l1, l2 elasticnet_0.1, ... elasticnet_0.9

로지스틱 회귀분석 모델을 정규화의 유형을 결정하기 위한 파라미터는 L1, L2, elasticnet로 지정하였다. 최적화 알고리즘을 반복해서 수행할 횟수는 3000으로 선택하였고, 각각의 파라미터에서 가중치를 최적화하는 알고리즘으로 L1은 liblinear을 사용하였고, L2는 lbfgs, elasticnet은 saga를 활용하였다. 각 포지션에 따라 각각 진행하였으며 5fold CV로 성능을 측정하였다. F1 score가 높아지도록, 전체 특징에서 하나씩 제거하는 방식으로 특징을 추출하였다. elasticnet의 Solver의 권장 사항(2014)에 따라 데이터 전처리를 standardscaler를 적용하였다. 예측 모델은 zscoring을 사용한 로지스틱 회귀분석 모델1, minmax를 사용한 로지스틱 회귀분석 모델2의 2가지 형태로 구분하여 나타냈다.

표 27. 로지스틱 회귀분석 모델 1 투수 및 포수 Hyper parameter

포지션	penalty	zscoring 변인
투수	elasticnet_0.1	win, lose, save, hold, 2B, IBB, LOB%, op_bat%, op_onbase%, P, IP/G, P/G, P/IP, CYP
포수	elasticnet_0.1	RBI, IBB, SF, IsoP, PSN, wRC, wRC/27, wRAA, wRC+

win: 승, lose:패, save: 세, hold: 홀드, 2B: 2루타, IBB: 고의 4구, LOB%: 잔루율, op_bat%: 상대타율, op_onbase%: 상대 출루율, P: 투구 수, IP/G: 경기당 이닝 수, P/G: 게임당 투구 수, P/IP: 이닝당 투구수, CYP: 사이영상 포인트, RBI: 타점, SF: 희비, IsoP: 순수장타율, PSN: 호타준족 점수, wRC: 조정 득점 창출력, wRC/27: 27아웃당 득점 생산력, wRAA: 리그평균대비 득점기여도, wRC+:파크팩터 적용 wRC

투수와 포수의 로지스틱 회귀분석 모델 구현 설정값은 동일하게 elasticnet_0.1을 조정하여 모형의 일반화 성능을 최적화하였다. 그에 따른 zscoring 변인 도출 결과 투수는 win, lose, save, hold, 2B, IBB, LOB%, op_bat%, op_onbase%, P, IP/G, P/G, P/IP, CYP가 선정되었으며, 포수의 경우 RBI, IBB, SF, IsoP, PSN, wRC, wRC/27, wRAA, wRC+가 변인으로 나타났다.

표 28. 로지스틱 회귀분석 모델 1 내야수 Hyper parameter

포지션	penalty	zscoring 변인
1루수	elasticnet_0.7	HR, wRAA, wRC+
2루수	elasticnet_0.1	WAR, 2B, SF, WPA, PA, HR%, BB%, BB/K, IsoP, Spd, wRC, wRC+
3루수	elasticnet_0.1	TB, HBP, SO, GIDP, SAC, SF, WPA, PA, K%, IsoP, IsoD, wOBA, wRC, wRC/27, wRAA, wRC+
유격수	elasticnet_0.1	runs, GIDP, SAC, AVG, WPA, IsoD, Spd, wRC, wRC+

HR: 홈런, wRAA: 리그평균대비 득점기여도, wRC+: 파크팩터 적용 wRC, WAR: 대체선수대비 승리기여도, 2B: 2루타, SF: 희비, WPA: 승리 확률 기여도, PA, 타석, HR%: 홈런율, BB%: 볼넷율, BB/K: 삼진당 볼넷 수, IsoP: 순수장타율, Spd: 스피드 스코어, wRC: 조정득점 창출력, TB: 총루타, HBP: 몸에 맞는볼, SO: 삼진, GIDP: 병살, SAC:희타, K%: 삼진율, Isod: 순수 출루율, wOBA: 가중 출루율, wRC/27: 27아웃당 조정 득점력, runs: 득점, AVG: 타율

2루수, 3루수, 유격수의 로지스틱 회귀분석 모델 구현 설정값은 동일하게 elasticnet_0.1을 조정하여 모형의 일반화 성능을 최적화하였다. 1루수는 elasticnet_0.7의 설정값을 조정하여 모형을 최적화하였다. zscoring 변인 도출 결과 1루수는 HR, wRAA, wRC+가 나타났다. 2루수는 WAR, 2B, SF, WPA, PA, HR%, BB%, BB/K, IsoP, Spd, wRC, wRC+로 확인되었으며. 3루수는 TB, HBP, SO, GIDP, SAC, SF, WPA, PA, K%, IsoP, IsoD, wOBA, wRC, wRC/27, wRAA, wRC+가 변인으로 선정 되었다. 유격수의 경우, runs, GIDP, SAC, AVG, WPA, IsoD, Spd, wRC, wRC+로 확인되었다.

표 29. 로지스틱 회귀분석 모델 1 외야수 및 지명타자 Hyper parameter

포지션	penalty	zscoring 변인
외야수	l2	WAR, RBI, BB, GIDP, WPA, IsoP, BABIP, Spd, wRAA, wRC
지명타자	elasticnet_0.2	WPA, K%, BB/K, BABIP, PSN, wRAA, wOBA

WAR: 대체선수 대비 승리기여도, RBI: 타점, BB: 볼넷, GIDP: 병살, WPA: 승리 확률 기여도, IsoP: 순수장타율 wRAA: 리그평균대비 득점기여도, BABIP: 인플레이 타구 타율, Spd: 스피드 스코어, K%: 삼진율, BB/K: 삼진당 볼넷 수, PSN: 호타준족 점수 wRC: 조정 득점 창출력, wOBA: 가중 출루율

외야수의 로지스틱 회귀분석 모델 구현은 L2를 통한 정규화 설정을 조정하였고 지명 타자는 elasticnet_0.2를 조정하여 일반화 성능을 최적화하였다. zscoring 변인 도출 결과, 외야수는 WAR, RBI, BB, GIDP, WPA, IsoP, BABIP, Spd, wRAA, wRC값이 선정되었으며, 지명타자는 WPA, K%, BB/K, BABIP, PSN, wRAA, wOBA가 변인으로 확인되었다.

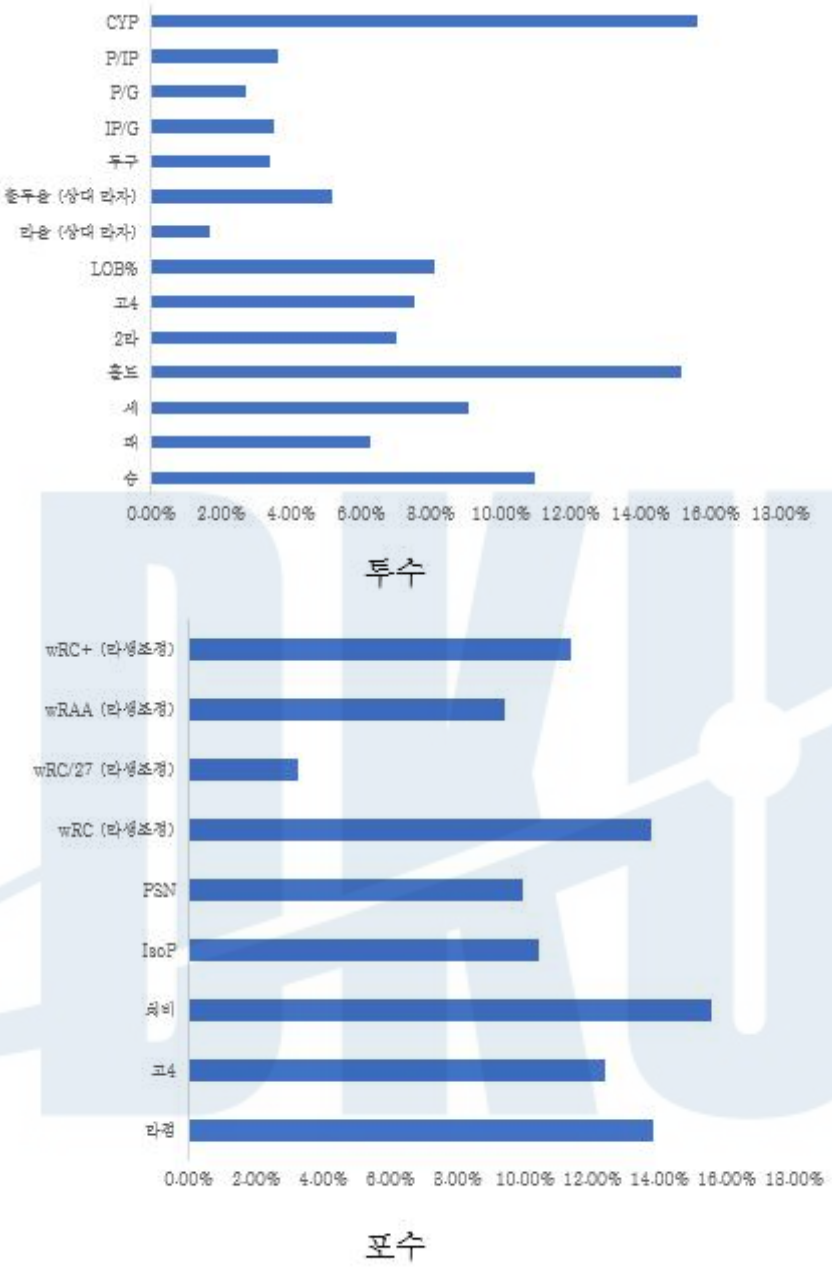
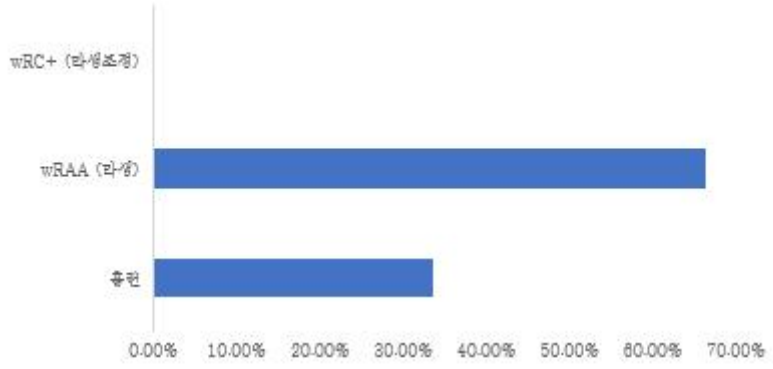
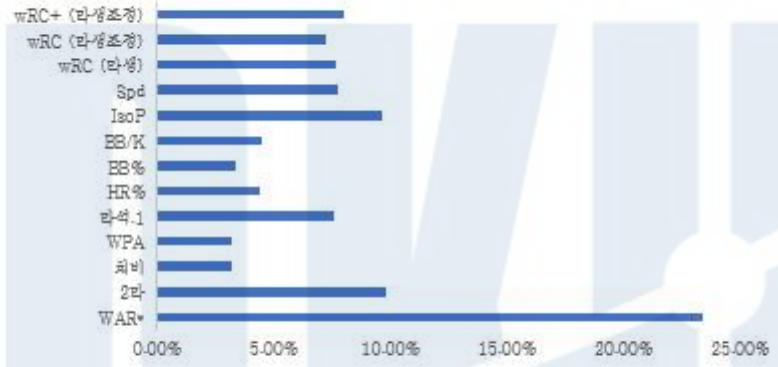


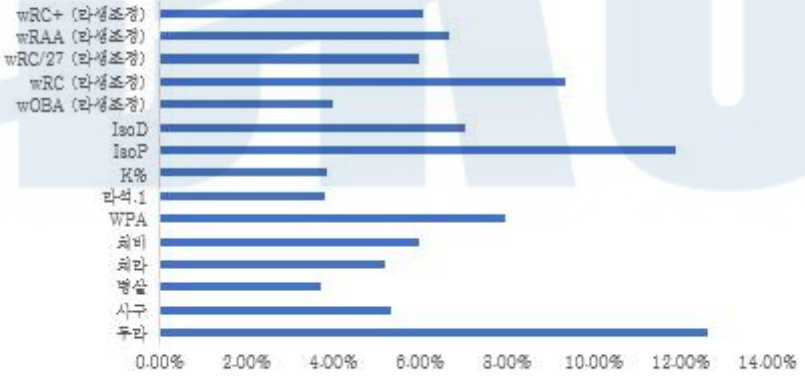
그림 7. 로지스틱 회귀분석 모델 1 투수, 포수 변인 중요도



1루수

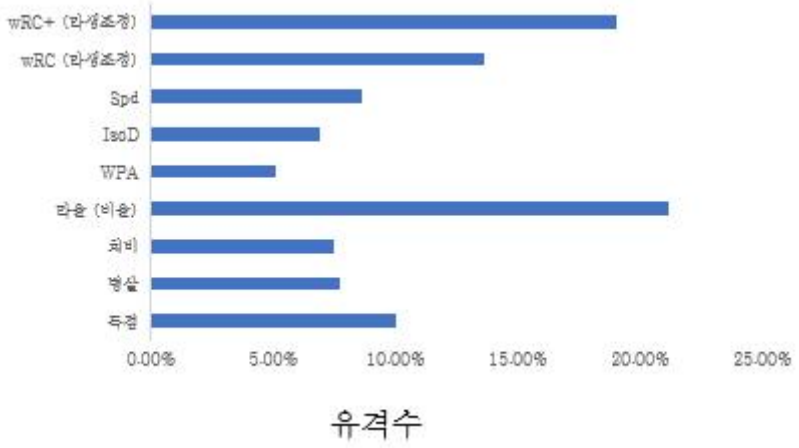


2루수



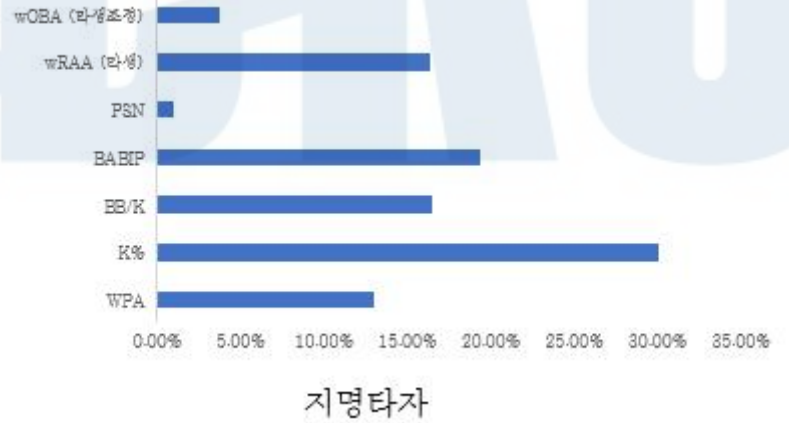
3루수

그림 8. 로지스틱 회귀분석 모델 1 1루수, 2루수, 3루수 변인 중요도



유격수

외야수



지명타자

그림 9. 로지스틱 회귀분석 모델 1 유격수, 외야수, 지명타자 변인 중요도

로지스틱 회귀분석 모델 1의 포지션별 변인 중요도를 <그림 7>, <그림 8>, <그림 9>과 같이 나타냈다. 투수의 경우 CYP가 15.65%로 가장 높은 중요도를 보였으며, 홀드 15.19%, 승 10.97% 순으로 확인되었다. 포수의 경우 희비, 타점, wRC, 고의 4구 순으로 나타났음을 알 수 있다. 1루수는 wRAA와 홈런만 사용되었으며, 2루수의 경우 WAR가 23.42%로 가장 높게 확인되었으며, 2루타, IsoP지표가 높은 중요도를 보였다. 3루수의 경우 루타, IsoP, wRC가 높게 나타났으며, 유격수는 타율, wRC+, wRC, 득점이 높게 나타났음을 알 수 있다. 외야수의 경우 WAR이 가장 높게 나타났으며 지명타자는 K%, BABIP, BB/K순으로 확인되었다.

표 30. 로지스틱 회귀분석 모델 2 투수 및 포수 Hyper parameter

포지션	penalty	minmax 변인
투수	elasticnet_0.2	win, save, LOB%, WHIP, WHIP+, P/G, P/IP, CYP
포수	elasticnet_0.2	runs, 2B, 3B, IBB, SAC, SF, WPA, PSN, wRC

win: 승, save: 세, LOB%: 잔루율, WHIP: 이닝당 출루허용율, WHIP+: 파크팩터 적용 이닝당 출루허용율, P/G: 게임당 투구수, P/IP: 이닝당 투구수, CYP: 사이영상 포인트 WPA: 승리 확률 기여도, wRC: 조정 득점 창출력, runs: 득점, 2B: 2루타, 3B: 3루타, IBB: 고의 4구, SAC: 희타, SF: 희비, PSN: 호타준족 점수

투수와 포수의 로지스틱 회귀분석 모델 구현 설정값은 동일하게 elasticnet_0.2를 조정하여 모형의 일반화 성능을 최적화하였다. 그에 따른 minmax 변인 도출 결과, 투수는 wwin, save, LOB%, WHIP, WHIP+, P/G, P/IP, CYP로 확인되었으며, 포수의 경우 runs, 2B, 3B, IBB, SAC, SF, WPA, PSN, wRC가 변인으로 나타났음을 알 수 있다.

표 31. 로지스틱 회귀분석 모델 2 내야수 Hyper parameter

포지션	penalty	minmax 변인
1루수	elasticnet_0.2	AB, SF, WPA, wRAA
2루수	elasticnet_0.5	WAR, runs, SO, BB%, wRAA, wRC+
3루수	elasticnet_0.3	HR%, wRC+
유격수	elasticnet_0.3	GIDP, SF, OBP, BB/K, IsoP, Spd, wOBA, wRC/27, wRAA

AB: 타수, SF: 희비, WPA: 승리 확률 기여도, wRAA: 리그평균대비 득점기여도, WAR: 대체선수 대비 승리기여도, runs: 득점, SO: 삼진, BB/%: 볼넷율, wRC+: 파크팩터 조정 wRC, HR%: 홈런율, GIDP: 병살, OBP: 출루율, BB/K: 삼진당 볼넷 수, Spd: 스피드 스코어 IsoP: 순수장타율 wRC/27: 27아웃당 득점 창충력 wRC: 조정 득점 창출력, wOBA: 가중 출루율

1루수의 로지스틱 회귀분석 모델 구현 설정값은 elasticnet_0.2을 조정하여 모형의 일반화 성능을 최적화하였다. minmax 변인 도출 결과 AB, SF, WPA, wRAA로 나타났으며, 2루수의 로지스틱 회귀분석 모델 구현 설정값은 elasticnet_0.5를 조정하였다. WAR, runs, SO, BB%, wRAA, wRC+로 확인되었다. 3루수, 유격수는 모델 구현 설정값을 동일하게 elasticnet_0.3을 조정하였다. minmax 변인 도출 결과 3루수는 HR%, wRC+임을 알 수 있었으며, 유격수는 GIDP, SF, OBP, BB/K, IsoP, Spd, wOBA, wRC/27, wRAA로 변인이 선정되었다.

표 32. 로지스틱 회귀분석 모델 2 외야수 및 지명타자 Hyper parameter

포지션	penalty	minmax 변인
외야수	elasticnet_0.1	runs, HR, BB, HBP, GIDP, AVG, IsoP, BABIP, Spd, wOBA
지명타자	elasticnet_0.1	HIT, IBB, OPS, HR%, K%, BB/K

runs: 득점, HR: 홈런, BB: 볼넷, HBP: 몸에 맞는볼, GIDP: 병살, AVG: 타율, IsoP: 순수장타율, BABIP: 인플레이 타구 타율, Spd: 스피드 스코어 wOBA: 가중 출루율, HIT: 안타, IBB: 고의 4구, OPS: 장타율+출루율, HR%: 홈런율, K%: 삼진율, BB/K: 삼진당 볼넷 수

외야수와 지명타자의 로지스틱 회귀분석 모델 구현 설정값은 동일하게 elasticnet_0.2을 조정하여 모형의 일반화 성능을 최적화하였다. minmax 변인 도출 결과 외야수는 runs, HR, BB, HBP, GIDP, AVG, IsoP, BABIP, Spd, wOBA로 확인되었으며, 지명타자는 HIT, IBB, OPS, HR%, K%, BB/K로 변인이 선정되었다.

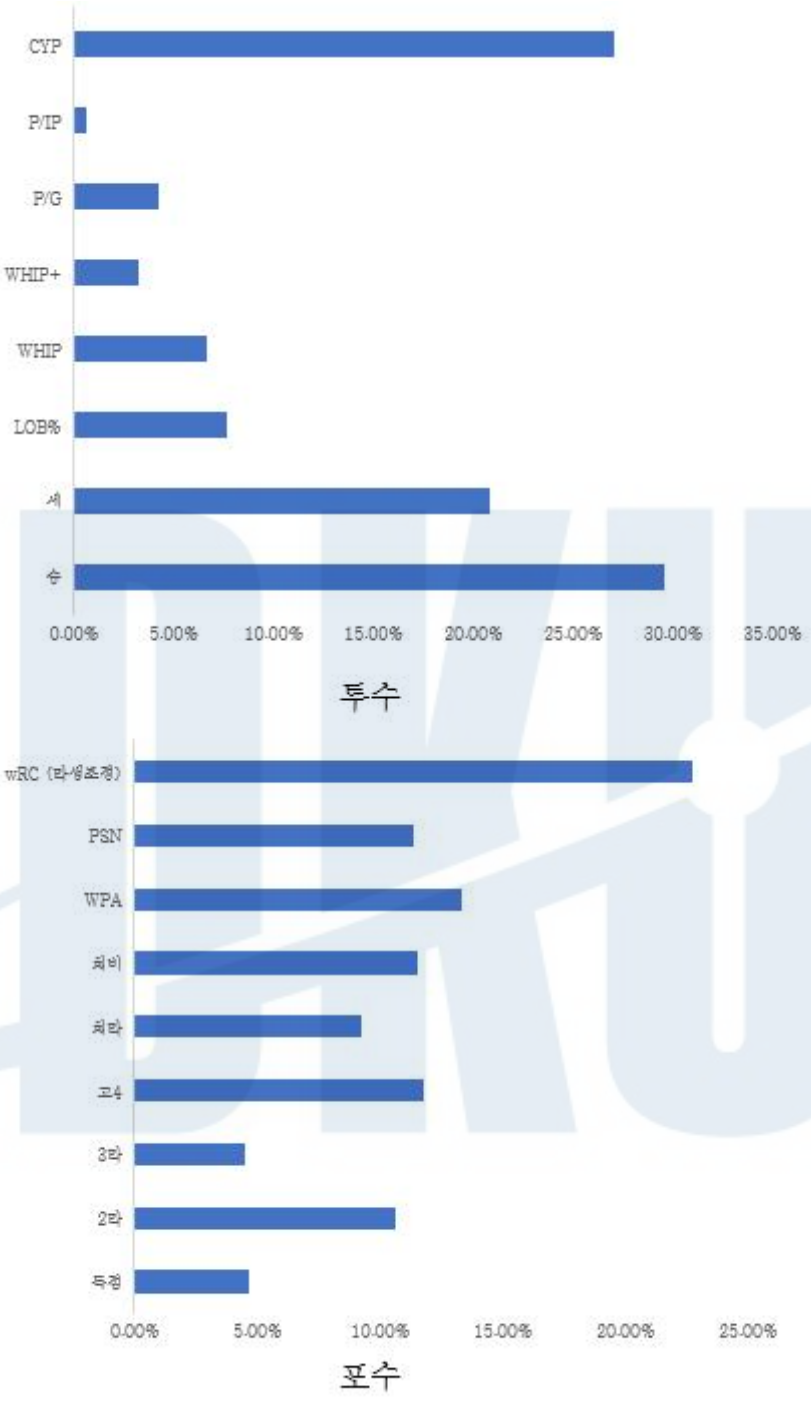


그림 10. 로지스틱 회귀분석 모델 2 투수, 포수 변인 중요도

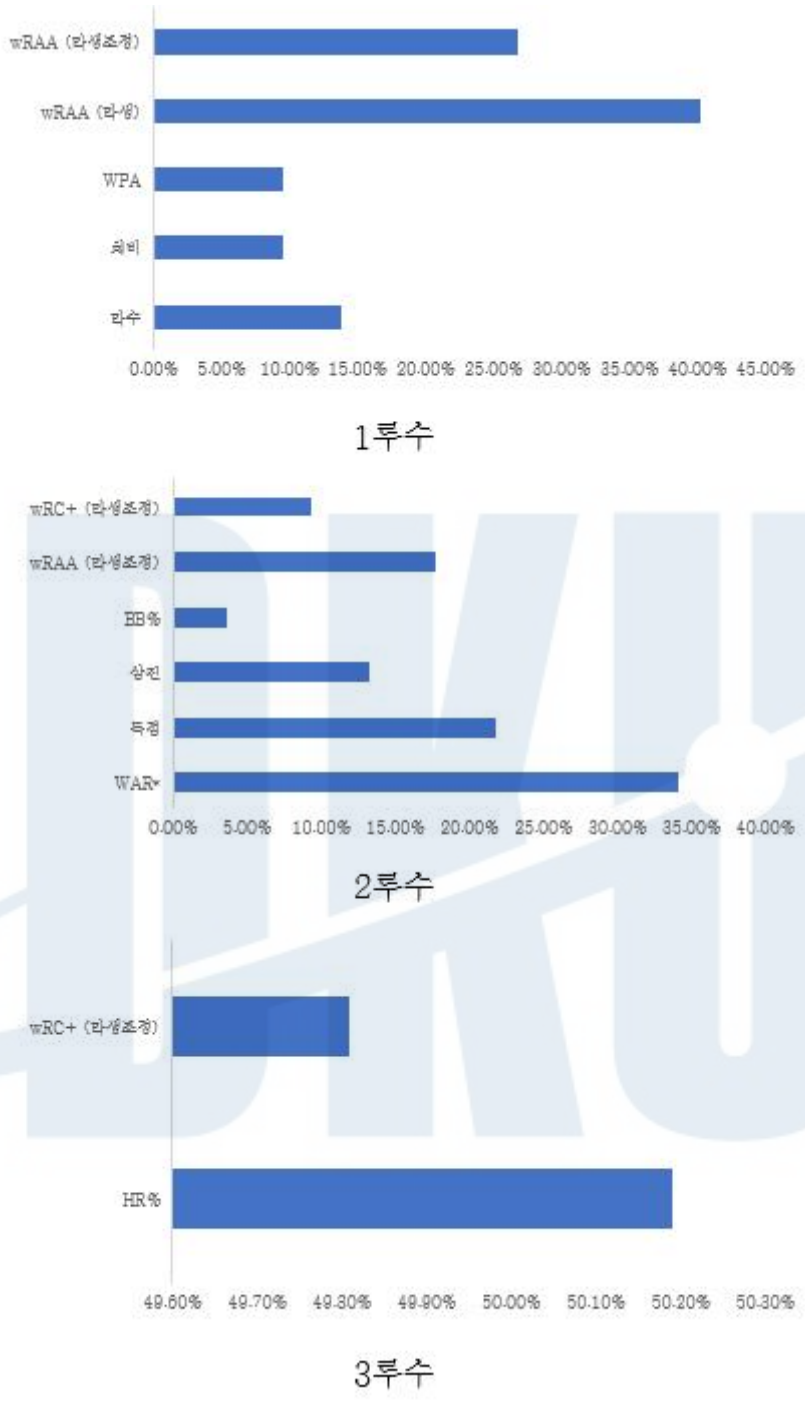


그림 11. 로지스틱 회귀분석 모델 2 1루수, 2루수, 3루수 변인 중요도

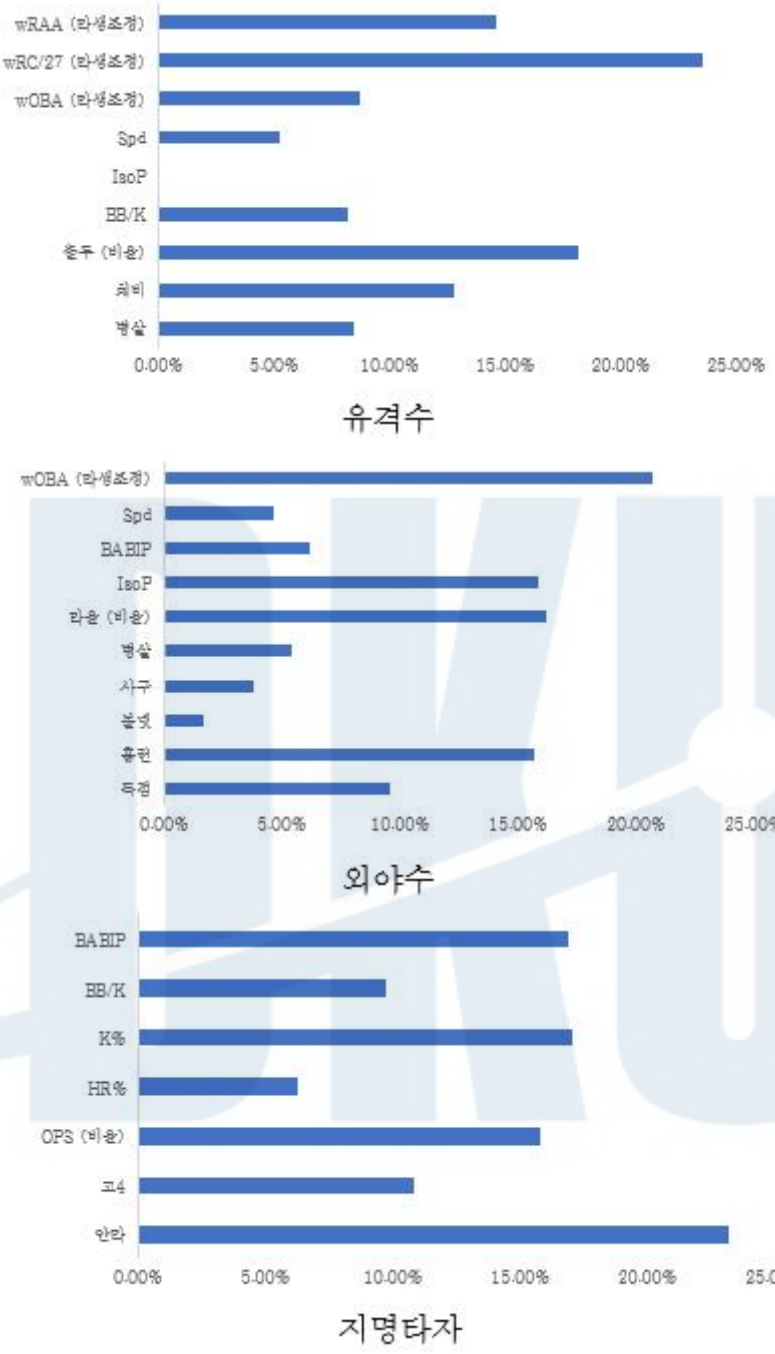


그림 12. 로지스틱 회귀분석 모델 2 유격수, 외야수, 지명타자 변인 중요도

로지스틱 회귀분석 모델 2의 포지션별 변인 중요도를 <그림 10>, <그림 11>, <그림 12>과 같이 구현하였다. 투수의 경우 승리 29.64%, CYP 27.05%, 세이브 20.85% 순으로 확인되었으며, 이닝당 투구수, WHIP+, 경기당 투구수가 0.68%, 3.21%, 4.25%로 상대적으로 낮은 중요도로 확인되었다. 포수의 경우 wRC가 22.75%로 가장 높은 중요도를 보였으며, WPA, 고의 4구, 희비, PSN 순으로 나타났다. 1루수의 경우 wRAA(타격생산력)이 40.24%로 가장 중요한 변인으로 확인되었으며, WPA가 9.5%로 상대적으로 낮은 중요도임을 확인할 수 있었다. 2루수의 경우 WAR이 가장 높게 나타났으며, BB%가 상대적으로 낮게 나타났고, 3루수의 경우 HR%와 wRC+ 두 지표가 변인으로 추출되었는데 50.19%, 49.81%로 확인되었다. 유격수의 경우 wRC/27, 출루율, wRAA, 희비 순으로 확인되었다. 외야수의 경우 wOBA, 타율, IsoP, 홈런 순으로 각각 20.71%, 16.19%, 15.93%, 15.74%로 나타났다. 지명타자에서는 안타 23.19%, K% 17.11%, BABIP 16.91%, OPS 15.85% 순으로 확인되었으며, HR%가 6.31%로 상대적으로 낮은 중요도로 나타났다.

(2) 서포트 백터 머신(Support vector machine)

표 33. 서포트 백터 머신 파라미터 설명

parameter	value
kernel	rbf, poly
degree	1,2,3
max_iter	5000
learning_rate	0.2
kind	rbf, poly_1, ..., poly_3

서포트 백터머신 모델을 최적화하기 위해 대상 Kernel은 모든 차수의 모든 다항식을 고려하여 고차원 차수에 매핑 될 수 있도록 하는 가우시안 방사 기저함수 rbf(radial basis function)와 복잡한 데이터를 더 높은 차원으로 변형하는 다항식 커널 Poly(Polynominal Kernel) 1, 2, 3으로 지정하였다. 각 포지션에 따라 각각 진행하였으며 5fold CV로 성능을 측정하였다. F1 score가 높아지도록, 전체 특징에서 하나씩 제거하는 방식으로 특징을 추출하였다. 예측 모델은 zscoring을 사용한 서포트 백터 머신 모델 1과 minmax를 사용한 서포트 백터머신 모델 2의 2가지 형태로 나타냈다.

표 34. 서포트 백터 머신 모델 1 투수 및 포수 Hyper parameter

포지션	kernel	zscoring 변인
투수	poly_3	WAR, 2B, WP, HR/9, PFR, op_onbase%, OPS, WHIP+, P, P/IP, CYP
포수	poly_1	AB, SB, IBB, SF, BABIP, PSN, wRC/27, wOBA, wRC, wRAA, wRC+

WAR: 대체선수 대비 승리기여도, 2B: 2루타, WP: 폭투, HR/9: 9이닝당 홈런 수, PFR: (삼진+볼넷)/이닝, op_onbase%: 상대 출루율, OPS: 장타율+출루율 WHIP+: 파크팩터 조정 이닝당 출루허용율, P: 투구수, P/IP: 이닝당 투구수, CYP: 사이영상 포인트, AB: 타수, SB: 도루, IBB: 고의 4구, SF: 희비, BABIP: 인플레이 타구 타율, PSN: 호타준족 점수 wRC: 조정 득점 창출력, wOBA: 가중 출루율, wRAA: 리그평균대비 득점기여도, wRC/27: 27아웃당 득점 창출력

투수는 서포트 벡터머신 모델 구현 설정은 poly_3를 조정하여 모형의 일반화 성능을 최적화하였다. zscoring 변인 도출 결과, WAR, 2B, WP, HR/9, PFR, op_onbase%, OPS, WHIP+, P, P/IP, CYP로 확인되었다. 포수는 서포트 백터 머신 모델 구현 설정값을 poly_1을 조정하여 모형의 일반화 성능을 최적화하였다. zscoring 변인 도출 결과, AB, SB, IBB, SF, BABIP, PSN, wRC/27, wOBA, wRC, wRAA, wRC+임을 알 수 있었다.

표 35. 서포트 백터 머신 모델 1 내야수 Hyper parameter

포지션	kernel	zscoring 변인
1루수	rbf	runs, 3B, TB, CS, HBP, IBB, GIDP, SLG, WPA, PA, HR%, IsoP, IsoD, BABIP, wRC, wRc/27, wRAA, wOBA, wRC+
2루수	poly_1	PA, BB/K, IsoP, IsoD, Spd, wRAA, wRC+
3루수	rbf	BB%, IsoP, PSN, wRC, wRC/27
유격수	poly_1	TB, RBI, CS, GIDP, SAC, OBP, PA, BB%, K%, BB/K, IsoD, BABIP, wRC, wRC/27, wRAA, wRC+

runs: 득점, 3B: 3루타, TB: 총 루타, CS: 도실, HBP: 몸에 맞는볼, GIDP: 병살, SLG: 장타율, PA: 타석, HR%: 홈런율, BB/K: 삼진당 볼넷 수, Spd: 스피드 스코어, PSN: 호타준족 점수 , BB%: 볼넷율, RBI: 타점, SAC: 희타, OBP: 출루율, ・K%: 삼진율 IsoP: 순수 장타율, IsoD: 순수 출루율, BABIP: 인플레이 타구 타율, wRC: 조정 득점 창출력, wOBA: 가중 출루율, wRAA: 리그평균대비 득점기여도, wRC/27: 27아웃당 득점 창출력, WPA: 승리확률 기여도, wRC+: 파크팩터 적용 wRC

1루수, 3루수는 서포트 벡터머신 모델 구현 설정은 rbf를 조정하여 모형의 일반화 성능을 최적화하였다. zscoring 변인 도출 결과, 1루수는 runs, 3B, TB, CS, HBP, IBB, GIDP, SLG, WPA, PA, HR%, IsoP, IsoD, BABIP, wRC, wRc/27, wRAA, wOBA, wRC+로 확인되었으며, 3루수는 BB%, IsoP, PSN, wRC, wRC/27이 변인으로 탐색되었다. 2루수, 유격수의 서포트 벡터머신 모델 구현 설정값은 poly_1을 조정하여 모

형의 일반화 성능을 최적화하였다. zscoring 변인 도출 결과, 2루수는 PA, BB/K, IsoP, IsoD, Spd, wRAA, wRC+가 나타났으며, 유격수는TB, RBI, CS, GIDP, SAC, OBP, PA, BB%, K%, BB/K, IsoD, BABIP, wRC, wRC/27, wRAA, wRC+가 변인으로 확인되었다.

표 36. 서포트 백터 머신 모델 1 외야수 및 지명타자 Hyper parameter

포지션	kernel	zscoring 변인
외야수	rbf	runs, TB, HBP, AVG, WPA, BABIP, wRC
지명타자	poly_1	HIT, IBB, SAC, SF, AVG, OBP, OPS, wOBA, PA, HR%, BB%, K%, BB/K, IsoP, IsoD, Spd, PSN, wRC, wRC/27, wRAA

runs: 득점, TB: 총루타, HBP: 몸에 맞는볼, AVG: 타율, WPA: 승리확률 기여도, WPA: 승리 확률 기여도, wRC: 조정 득점 창출력, BABIP: 인플레이 타구 타율, HIT: 안타, IBB: 고의 4구, SAC: 희타, SF: 희비, OBP: 출루율, OPS: 장타율+출루율, PA: 타수, wOBA: 가중 출루율, HR%: 홈런율, BB%: 볼넷율, K%: 삼진율, BB/K: 삼진당 볼넷 수, Spd: 스피드 스코어, PSN: 호타준족 점수 IsoP: 순수 장타율, IsoD: 순수 출루율, wRAA: 리그평균대비 득점기여도, wRC/27: 27아웃당 득점 창출력

외야수는 서포트 벡터머신 모델 구현 설정은 rbf를 조정하여 모형의 일반화 성능을 최적화하였다. zscoring 변인 도출 결과, runs, TB, HBP, AVG, WPA, BABIP, wRC로 선정되었다. 지명타자는 모델 구현 설정값을 poly_1을 조정하여 일반화 성능을 최적화하였다. zscoring 변인 결과, HIT, IBB, SAC, SF, AVG, OBP, OPS, wOBA, PA, HR%, BB%, K%, BB/K, IsoP, IsoD, Spd, PSN, wRC, wRC/27, wRAA가 변인임을 알 수 있다.

표 37. 서포트 백터 머신 모델 2 투수 및 포수 Hyper parameter

포지션	kernel	minmax 변인
투수	poly_2	win, WP, BB%, LOB%, op_onbase%, P/IP, CYP
포수	poly_3	AB, RBI, HBP, IBB, SAC, AVG, wOBA, WPA, BB%, IsoP, IsoD, BABIP, Spd, PSN, wRC, wRC/27, wRAA, wRC+

win:승, WP: 폭투, BB%: 볼넷율, LOB%: 잔루율, op_onbase%: 상대 출루율, P/IP: 이닝당 투구수, CYP: 사이영상 포인트, AB: 타수, RBI: 타점, HBP: 몸에 맞는볼, IBB: 고의 4구, SAC: 희타, AVG: 타율, wOBA: 가중 출루율, WPA: 승리 확률 기여도, wRC: 조정 득점 창출력, BABIP: 인플레이 타구 타율, wOBA: 가중 출루율, BB%: 볼넷율, BB/K: 삼진당 볼넷 수, Spd: 스피드 스코어, PSN: 호타준족 점수 IsoP: 순수 장타율, IsoD: 순수 출루율, wRAA: 리그평균대비 득점기여도, wRC/27: 27아웃당 득점 창출력

투수의 서포트 벡터 머신의 구현 모델 설정은 poly_2를 조정하여 모형의 일반화 성능을 최적화하였다. minmax 변인 도출 결과, win, WP, BB%, LOB%, op_onbase%, P/IP, CYP로 도출되었다. 포수는 poly_3의 설정값을 조정하여 서포트 벡터 머신 모델 모형을 최적화하였다. minmax 변인 도출 결과, AB, RBI, HBP, IBB, SAC, AVG, wOBA, WPA, BB%, IsoP, IsoD, BABIP, Spd, PSN, wRC, wRC/27, wRAA, wRC+로 확인되었다.

표 38. 서포트 백터 머신 모델 2 내야수 Hyper parameter

포지션	kernel	minmax 변인
1루수	rbf	3B, GIDP, HR%, BB%, IsoP, IsoD, BABIP, PSN, wOBA, wRC, wRC/27, wRAA, wRC+
2루수	poly_3	runs, 3B, CS, HBP, IBB, GIDP, SF, HR%, BB/K, IsoP, IsoD, PSN, wRC, wRC+
3루수	poly_3	Game, 2B, TB, RBI, BB, SAC, HR%, BB%, IsoP, Spd, PSN, wRAA, wOBA, wRC, wRC/27, wRAA, wRC+
유격수	poly_2	Game, runs, 2B, HBP, IBB, GIDP, SAC, SF, AVG, BB%, BB/K, IsoP, Spd, wRC, wRC/27, wRC+

3B: 3루타, GIDP: 병살, HR%: 홈런율, BB%: 볼넷율, IsoP: 순수장타율, IsoD: 순수출루율, BABIP: 인플레이 타구 타율, PSN: 호타준족 점수, wOBA: 가중 출루율, wRC: 조정 득점 창출력, wRC/27: 27아웃당 득점 생산력, wRAA: 리그평균대비 득점기여도, wRC+: 파크팩터 적용 wRC, runs: 득점, CS: 도실, HBP: 몸에 맞는볼, IBB: 고의 4구, GIDP: 병살, SF: 희비, Spd: 스피드 스코어, Game: 출장, 2B: 2루타, SAC: 희타, AVG: 타율, BB/K: 삼진당 볼넷 수

1루수의 서포트 벡터 머신의 구현 모델 설정은 rbf를 조정하여 모형의 일반화 성능을 최적화하였다. minmax 변인 도출 결과, 3B, GIDP, HR%, BB%, IsoP, IsoD, BABIP, PSN, wOBA, wRC, wRC/27, wRAA, wRC+가 나타났다 2루수, 3루수는 poly_3의 설정값을 조정하여 서포트 벡터 머신 모델 모형을 최적화하였다. minmax 변인 도출 결과, 2루수는 runs, 3B, CS, HBP, IBB, GIDP, SF, HR%, BB/K, IsoP, IsoD, PSN, wRC, wRC+로 확인되었으며, 3루수는 Game, 2B, TB, RBI, BB, SAC, HR%, BB%, IsoP, Spd, PSN, wRAA, wOBA, wRC, wRC/27, wRAA, wRC+가 도출되었다. 유격수는 poly_2의 설정값을 조정하여 서포트 벡터머신 모델 모형을 최적화하였다. minmax 변인 도출 결과, Game, runs, 2B, HBP, IBB, GIDP, SAC, SF, AVG, BB%, BB/K, IsoP, Spd, wRC, wRC/27, wRC+로 나타났음을 알 수 있다.

표 39. 서포트 백터 머신 모델 2 외야수 및 지명타자 Hyper parameter

포지션	kernel	minmax 변인
외야수	poly_3	WAR, runs, 2B, 3B, HR, RBI, SB, CS, IBB, GIDP, SAC, AVG, WPA, PA, K/%, IsoP, BABIP, PSN
지명타자	poly_2	HIT, OPS, WPA, HR%, K%, IsoD, BABIP, PSN, wRAA, wRC+

WAR: 대체선수 대비 승리기여도, runs: 득점, 2B: 2루타, 3B: 3루타, HR: 홈런, RBI: 타점, SB: 도루, CS: 도실, IBB: 고의 4구, GIDP: 병살, SAC: 희타, AVG: 타율, WPA: 승리 확률 기여도, PA: 타석, K%: 삼진율, IsoP: 순수 장타율, BABIP: 인플레이 타구 타율, PSN: 호타준족 점수, HIT: 안타, OPS: 장타율+출루율, HR%: 홈런율, K%: 삼진율, IsoD: 순수 출루율, wRAA: 리그평균대비 득점기여도, wRC+: 파크팩터 조정 wRC

외야수의 서포트 벡터 머신의 구현 모델 설정은 poly_3를 조정하여 모형의 일반화 성능을 최적화하였다. minmax 변인 도출 결과, WAR, runs, 2B, 3B, HR, RBI, SB, CS, IBB, GIDP, SAC, AVG, WPA, PA, K/%, IsoP, BABIP, PSN가 변인으로 도출되었다. 지명타자는 poly_2의 설정값을 조정하여 서포트 벡터 머신 모델 모형을 최적화하였다. minmax 변인 도출 결과, HIT, OPS, WPA, HR%, K%, IsoD, BABIP, PSN, wRAA, wRC+로 나타났음을 확인할 수 있다.

(3) 랜덤포레스트(Random Forest)

표 40. 랜덤포레스트 파라미터 설명

parameter	value
n_estimators	25, 50, 100
criterion	gini, entropy
learning_rate	0.2
kind	25_gini, 25_entropy,... 100_entropy

랜덤포레스트 모델을 최적화하기 위해 대상 n_estimator는[25, 50, 100], 지식 노드에 있는 데이터가 얼마나 섞여 있는지를 나타내는 불순도 수치를 나타내는 함수로 [entropy, gini]로 지정하였다. 학습율은 0.2로 지정하였다. 각 포지션에 따라 각각 진행하였으며 5fold CV로 성능을 측정하였다. F1 score가 높아지도록, 전체 특징에서 하나씩 제거하는 방식으로 특징을 추출하였다. 예측 모델은 zscoring을 사용한 랜덤포레스트 모델1과 minmax를 사용한 랜덤포레스트 모델2의 2가지 형태로 나타냈다.

표 41. 랜덤포레스트 모델 1 투수 및 포수 Hyper parameter

포지션	criterion	zscoring 변인
투수	25_gini	WAR, game, CG, SO, start, win, lose, save, hold, IP, runs, ER, TA, HIT, 2B, 3B, HR, BB, IBB, HBP, SO, WP, ERA, FIP, WHIP, ERA+, FIP+, WPA, K/9, BB/9, K/BB, HR/9, K%, BB%, K-BB%, PFR, LOB%, op_bat%, op_onbase%, SLG, OPS, WHIP+, P, IP/G, P/G, P/IP, P/PA, CYP
포수	25_entropy	game, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC, wRC/27, wRAA, wOBA, wRC+

투수의 랜덤포레스트 구현 모델 설정값은 25_gini를 조정하여 모형의 일반화 성능을 최적화하였다. zscoring 변인 도출 결과, WAR, game, CG, SO, start, win, lose, save, hold, IP, runs, ER, TA, HIT, 2B, 3B, HR, BB, IBB, HBP, SO, WP, ERA, FIP, WHIP, ERA+, FIP+, WPA, K/9, BB/9, K/BB, HR/9, K%, BB%, K-BB%, PFR, LOB%, op_bat%, op_onbase%, SLG, OPS, WHIP+, P, IP/G, P/G, P/IP, P/PA, CYP로 대부분의 지표가 변인으로 도출되었다. 포수는 25_entropy의 설정값을 조정하여 랜덤포레스트 모델 모형을 최적화하였다. zscoring 변인 도출 결과, game, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC, wRC/27, wRAA, wOBA, wRC+로 확인되었다.

표 42. 랜덤포레스트 모델 1 내야수 Hyper parameter

포지션	criterion	zscoring 변인
1루수	25_entropy	WAR, game, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, WPA, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC, wRC/27, wRAA, wRC+,
2루수	25_gini	WAR, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, OPS, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC/27, wRAA, wOBA , wRC , wRC+
3루수	25_gini	WAR, G, PA, AB, runs, HIT, 2B, 3B, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC, wRC/27, wRAA, wRC+
유격수	25_entropy	WAR, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC, wRC/27, wRAA wRC+

1루수의 랜덤포레스트 구현 모델 설정값은 25_entropy를 조정하여 모형의 일반화 성능을 최적화하였다. zscoring 변인 도출 결과, WAR, game, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, WPA, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC, wRC/27, wRAA, wRC+로 확인되었다. 2루수, 3루수는 25_gini 설정값을 조정하여 서포트 벡터 머신 모델 모형을 최적화하였다. zscoring 변인 도출 결과, 2루수는 WAR, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG,

OBP, OPS, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC/27, wRAA, wOBA , wRC , wRC+로 도출되었으며, 3루수의 경우, WAR, G, PA, AB, runs, HIT, 2B, 3B, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC, wRC/27, wRAA, wRC+로 나타났음을 알 수 있다. 유격수는 1루수와 동일한 랜덤포레스트 구현 모델 설정값을 25_entropy를 조정하여 모형의 일반화 성능을 최적화하였다. zscoring 변인 도출 결과, WAR, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC, wRC/27, wRAA wRC+로 추출되었다.

표 43. 랜덤포레스트 모델 1 외야수 및 지명타자 Hyper parameter

포지션	criterion	zscoring값 변인
외야수	25_gini	WAR, PA, AB, runs, HIT, 2B, HR, TB, SB, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wRC/27, wRAA, wRC
지명타자	50_entropy	WAR, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, OPS, wOBA, wRC+, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wRC/27, wRAA, wOBA, wRC+

외야수의 랜덤포레스트 구현 모델 1 설정값은 25_gini를 조정하여 모형의 일반화 성능을 최적화하였다. zscoring 변인 도출 결과, WAR, PA, AB, runs, HIT, 2B, HR, TB, SB, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+,

WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wRC/27, wRAA, wRC가 나타났다. 포수는 25_entropy의 설정값을 조정하여 랜덤포레스트 모델 1 모형을 최적화하였다. zscoring 변인 도출 결과, gWAR, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, OPS, wOBA, wRC+, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wRC/27, wRAA, wOBA, wRC+로 확인되었다.

랜덤포레스트 모델 1의 포지션별 중요 변인을 <그림 13>, <그림 14>, <그림 15>과 같이 구현하였다. 랜덤포레스트모델 1에서는 많은 변인들이 추출되어 상대적으로 중요한 변인 15개를 추출하여 그래프로 나타냈다. 투수의 경우 승리, CYP, WAR, 출장 수 순으로 높은 경향을 보였다. 포수의 경우 wRC, wRC/27이 가장 높은 중요도로 확인되었다. 1루수의 경우 wRAA, wRC, WAR등 세이버 메트릭스 지표가 높은 중요도임을 알 수 있었으며, 2루수는 wRC, WAR, 득점, wRC+순으로 중요도가 높게 나타났다. 3루수는 HR%, wRC, IsoP가 높은 중요도를 보였으며, 유격수는 출루율, 타율, wRC/27, wRAA 순으로 확인되었다. 외야수의 경우 루타 수, WAR, wRC가 12.52%, 10.98%, 7.89%로 확인되었으며 지명타자는 타율 18.59%로 가장 높게 나타났으며, 안타 7.78%, wRC+ 4.91% 순으로 높게 나타났다.

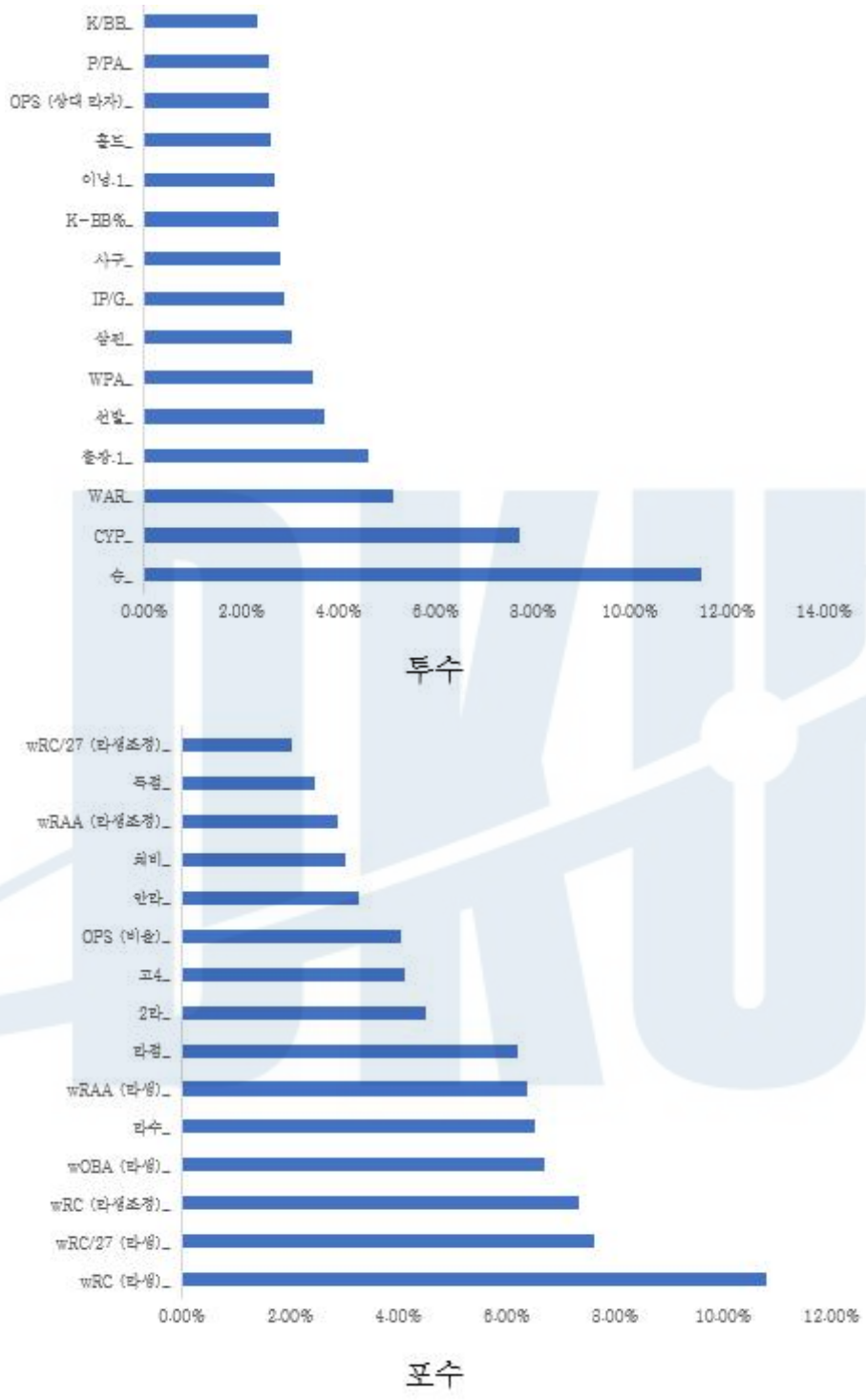


그림 13. 랜덤포레스트 모델 1 투수, 포수 변인 중요도

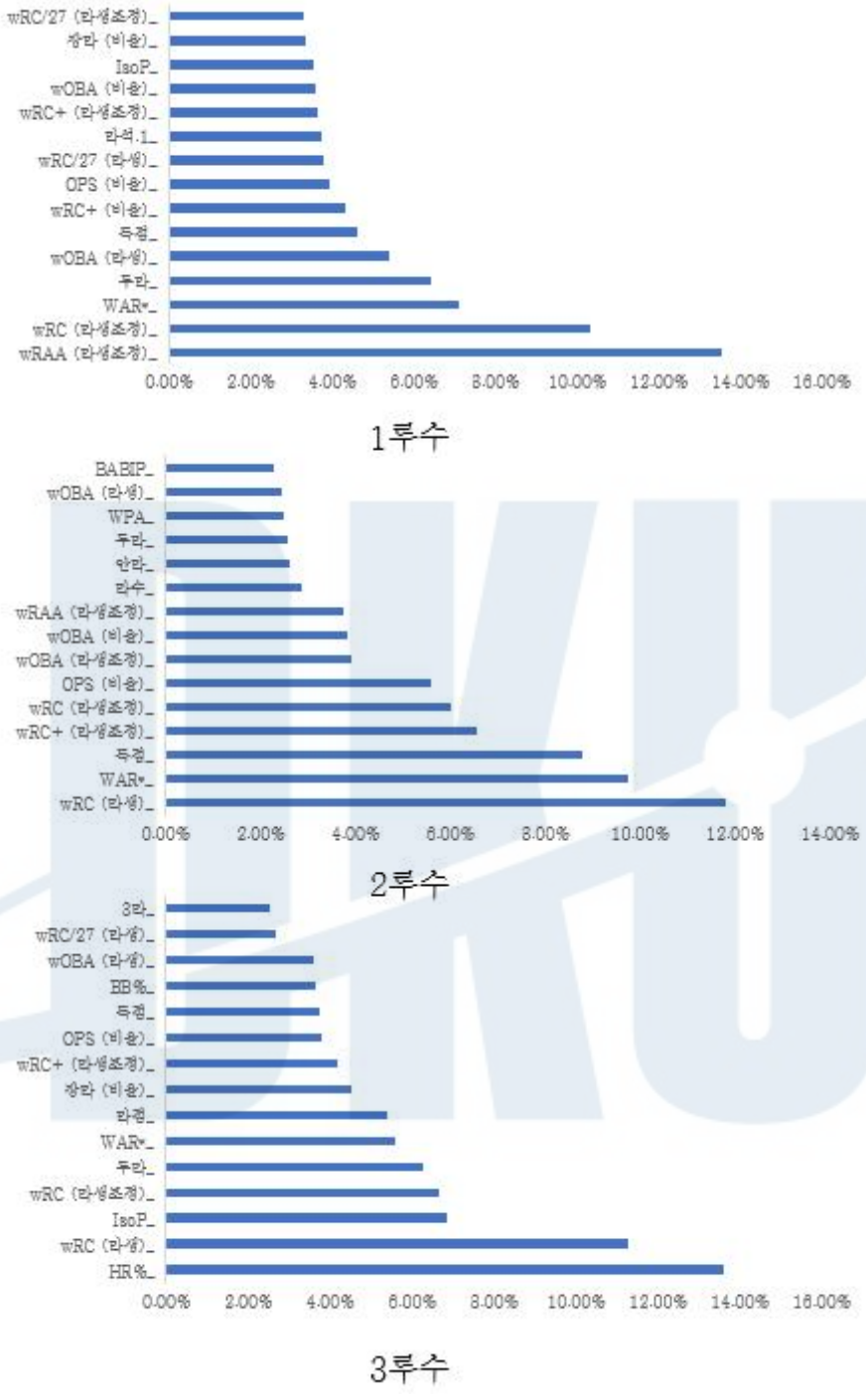


그림 14. 랜덤포레스트 모델 1 1루수, 2루수, 3루수 변인 중요도

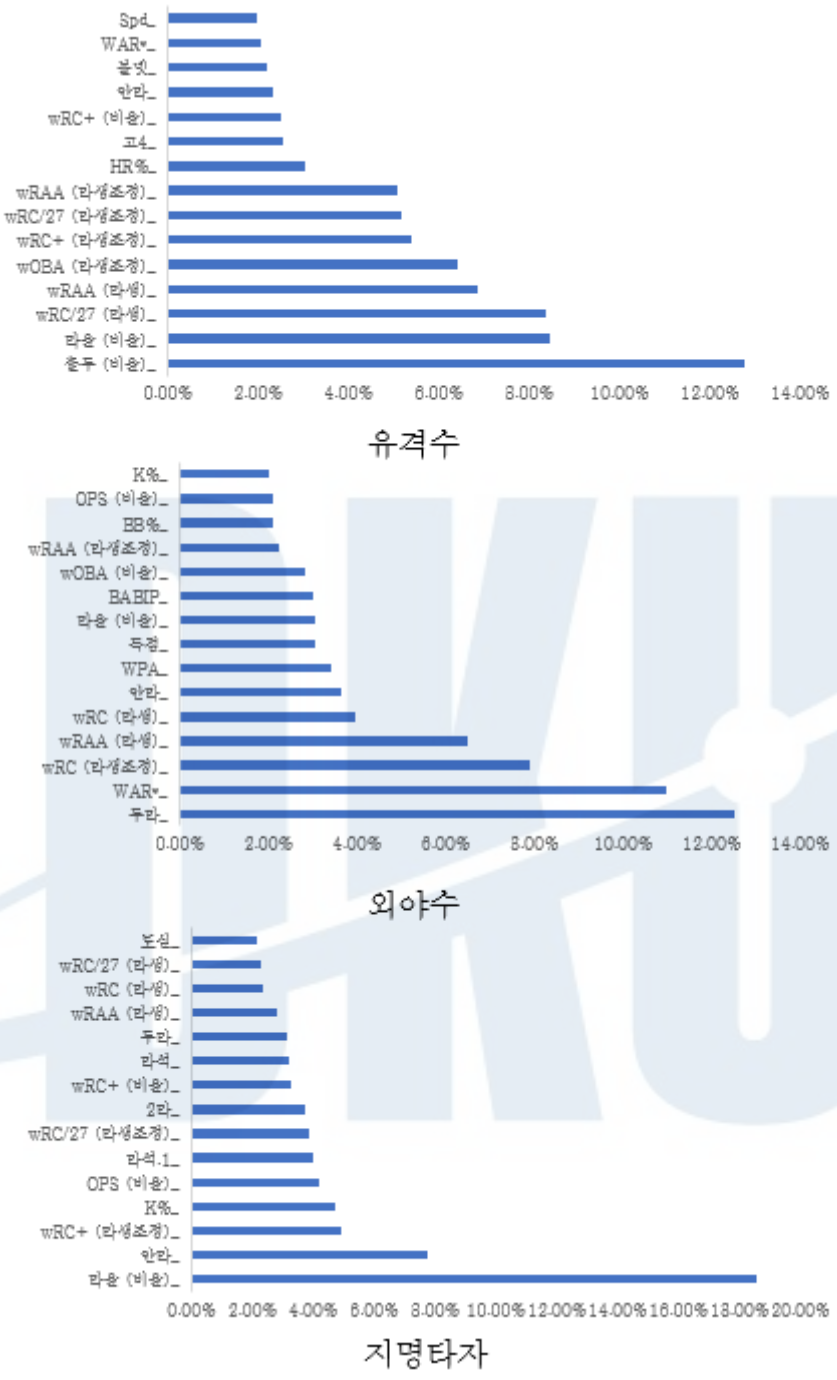


그림 15. 랜덤포레스트 모델 1 유격수, 외야수, 지명타자 변인 중요도

표 44. 랜덤포레스트 모델 2 투수 및 포수 Hyper parameter

포지션	criterion	minmax 변인
투수	25_gini	WAR, game, CG, SO, start, win, lose, save, hold, IP, runs, ER, batter, HIT, 2B, 3B, HR, BB, IBB, SO, WP, ERA, FIP, WHIP, ERA+, FIP+, WPA, K/9, BB/9, K/BB, HR/9, K%, BB%, PFR, BABIP, LOB%, op_bat%, op_onbase%, SLG, OPS, WHIP, WHIP+, P, IP/G, P/G, P/IP, P/PA, CYP
포수	100_gini	WAR, G, PA, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, IBB, SO, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wRC, wRC/27, wRAA

투수의 랜덤포레스트 구현 모델 설정값은 25_gini를 조정하여 모형의 일반화 성능을 최적화하였다. minmax 변인 도출 결과, WAR, game, CG, SO, start, win, lose, save, hold, IP, runs, ER, batter, HIT, 2B, 3B, HR, BB, IBB, SO, WP, ERA, FIP, WHIP, ERA+, FIP+, WPA, K/9, BB/9, K/BB, HR/9, K%, BB%, PFR, BABIP, LOB%, op_bat%, op_onbase%, SLG, OPS, WHIP, WHIP+, P, IP/G, P/G, P/IP, P/PA, CYP로 확인되었다. 포수는 100_gini의 설정값을 조정하여 랜덤포레스트 모델 2 모형을 최적화하였다. minmax 변인 도출 결과, WAR, G, PA, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, IBB, SO, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wRC, wRC/27, wRAA로 나타났음을 알 수 있다.

표 45. 랜덤포레스트 모델 2 내야수 Hyper parameter

포지션	criterion	minmax 변인
1루수	25_gini	WAR, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wRC/27, wRAA, wRC, wRAA
2루수	25_gini	WAR, G, PA, AB, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wRC, wRC/27, wRAA
3루수	50_entropy	HIT, 2B, TB, BB, HBP, IBB, SO, GIDP, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, PSN, wOBA , wRC/27 , wRAA , wRC+
유격수	25_entropy	WAR*, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC, wRAA, wRC/27, wRC+

1루수, 2루수의 랜덤포레스트 구현 모델 설정값은 25_gini를 조정하여 모형의 일반화 성능을 최적화하였다. minmax 변인 도출 결과, 1루수는 WAR, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wRC/27, wRAA, wRC, wRAA가 변인으로 추출되었으며, 2루수는 WAR, G, PA, AB, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wRC, wRC/27, wRAA로 확인되었다. 3루수는 50_entropy 설정값을 조정하여 랜덤포레스트 모델 2 모형을 최적화하였다. minmax 변인 도출 결과, HIT, 2B, TB, BB, HBP, IBB, SO, GIDP, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, BABIP, PSN, wOBA , wRC/27 , wRAA , wRC+가 나타났다 유격수는 25_entropy 설정값을 조정하여 랜덤포레스트 모델 모형을 최적화하였다. minmax 변인 도출 결과, WAR, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, K%, BB/K, IsoP, IsoD, BABIP, Spd, PSN, wOBA, wRC, wRAA, wRC/27, wRC+ 가 변인으로 나타났다.

표 46. 랜덤포레스트 모델 2 외야수 및 지명타자 Hyper parameter

포지션	criterion	minmax 변인
외야수	25_gini	WAR, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, Spd, PSN, wRC, wRC/27, wRAA
지명타자	50_entropy	WAR, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, Spd, PSN, wRC, wRC/27, wRAA

외야수의 랜덤포레스트 구현 모델 설정값은 25_gini를 조정하여 모형의 일반화 성능을 최적화하였다. minmax 변인 도출 결과, WAR, G, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, Spd, PSN, wRC, wRC/27, wRAA로 확인되었다. 지명타자는 50_entropy의 설정값을 조정하여 랜덤포레스트 모델 2 모형을 최적화하였다. minmax 변인 도출 결과, WAR, PA, AB, runs, HIT, 2B, 3B, HR, TB, RBI, SB, CS, BB, HBP, IBB, SO, GIDP, SAC, SF, AVG, OBP, SLG, OPS, wOBA, wRC+, WPA, HR%, BB%, K%, BB/K, IsoP, IsoD, Spd, PSN, wRC, wRC/27, wRAA가 변인으로 추출되었다.

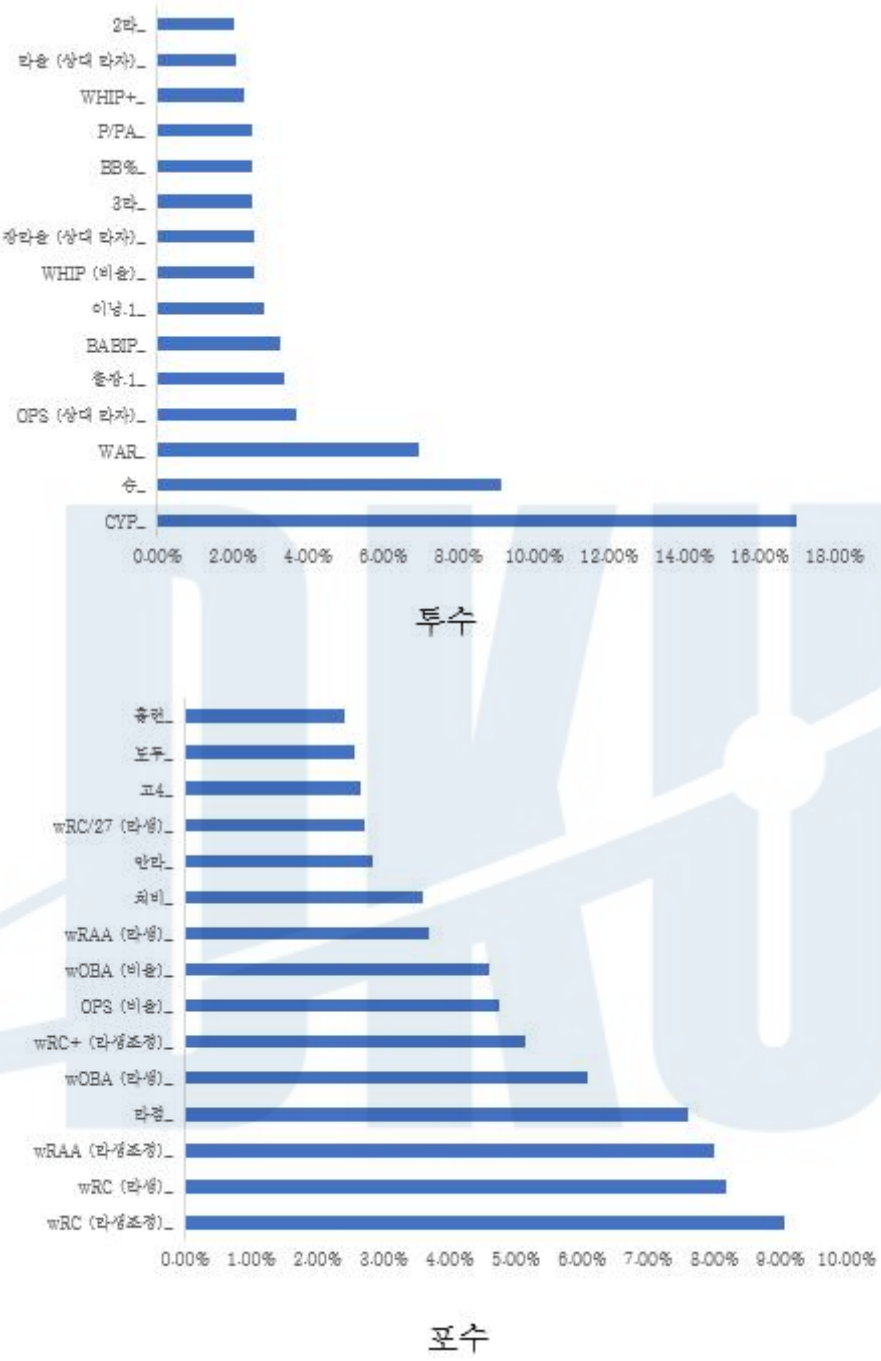


그림 16. 랜덤포레스트 모델 2 투수, 포수 변인 중요도

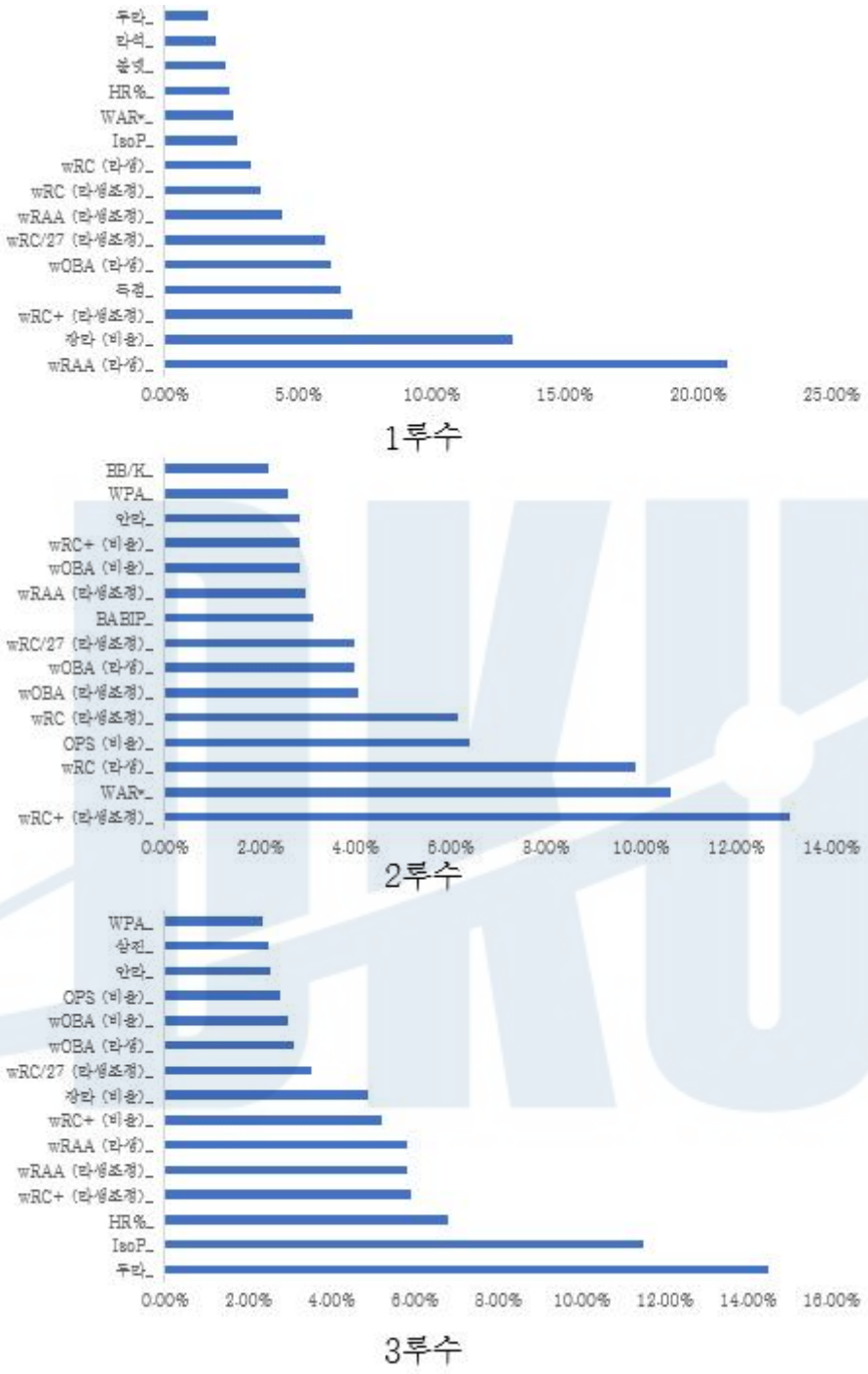


그림 17. 랜덤포레스트 모델 2 1루수, 2루수, 3루수 변인 중요도

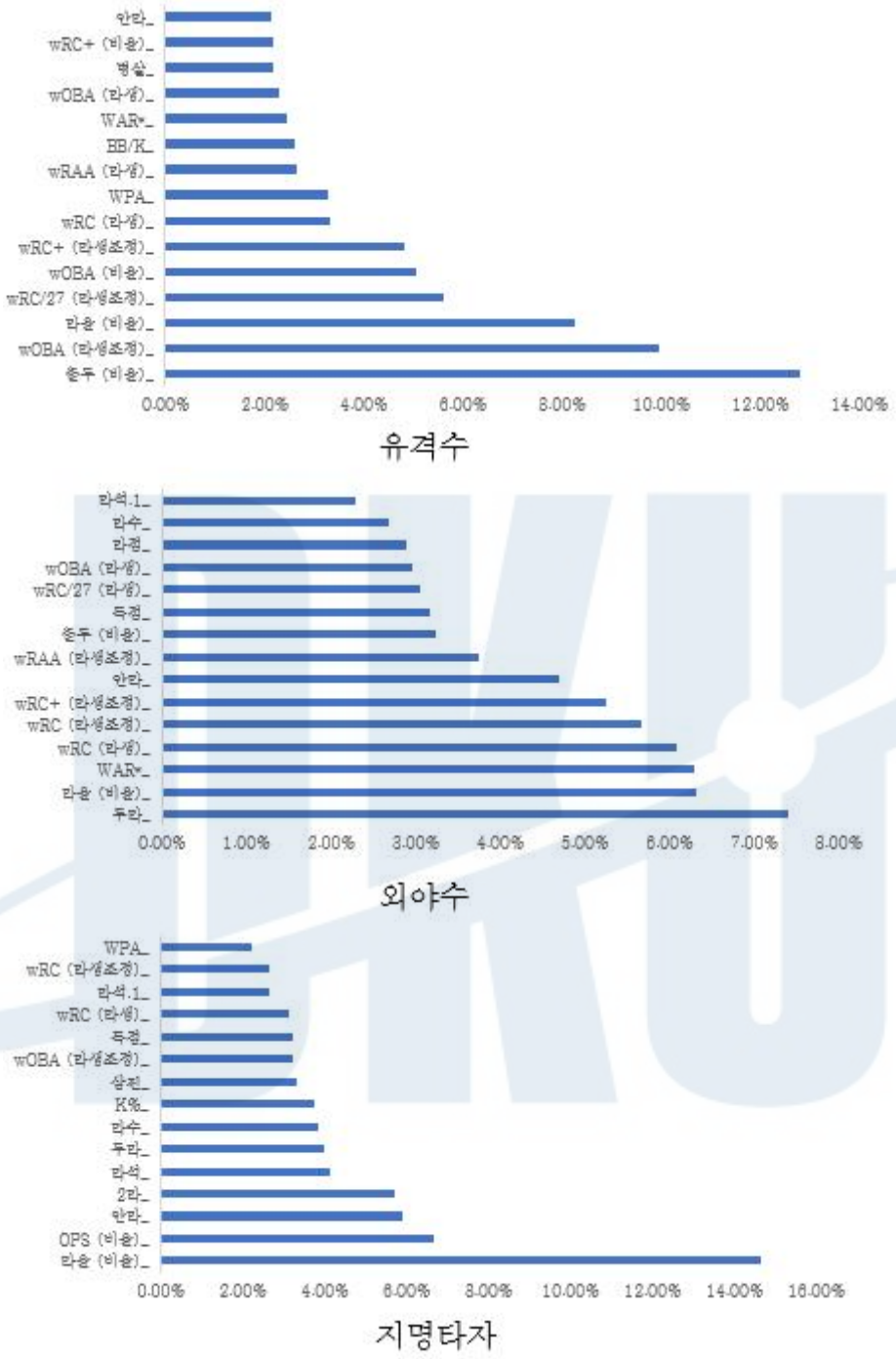


그림 18. 랜덤포레스트 모델 2 유격수, 외야수, 지명타자 변인 중요도

랜덤포레스트 모델 2의 포지션별 중요 변인을 <그림 16>, <그림 17>, <그림 18>와 같이 구현하였다. 랜덤포레스트모델 2에서는 많은 변인들이 추출되어 상대적으로 중요한 변인 15개를 추출하여 그래프로 나타냈다. 투수의 중요 변인으로는 CYP가 16.97%로 가장 높게 나타났으며, 승리 수, WAR, 상대타자 OPS 순으로 높게 나타났다. 포수의 경우 조정 wRC, wRC, wRAA, 타점, wOBA 순으로 중요 변인이 추출되었다. 1루수는 wRAA가 21.07%로 가장 중요한 변인으로 확인되었으며, 장타율, wRC+, 득점, wOBA 순으로 추출되었다. 2루수는 wRC+, WAR이 가장 중요한 변인으로 추출되었으며, wRC, OPS 순으로 나타났다. 3루수는 루타 수가 가장 높은 14.49%로 나타났으며 IsoP, HR%, wRC+순으로 중요 변인이 추출되었다. 유격수의 경우 출루율이 가장 중요한 변인으로 추출되었으며 wOBA, 타율, wRC/27 순으로 중요 변인이 확인되었다. 외야수는 루타 수가 7.40%로 가장 중요한 변인으로 나타났으며, 타율 6.32%, WAR 6.29%, wRC 6.09% 순으로 중요 변인이 추출되었다. 지명타자는 타율이 14.66%로 가장 중요한 변인으로 확인되었으며 OPS, 안타, 2루타, 타석 순으로 중요 변인이 나타났다.

(4) XG부스트(XGBoost)

표 47. XGBoost 파라미터 설명

parameter	value
n_estimators	25, 50, 100
criterion	exact, approx, hist
learning_rate	0.2
kind	25_exact,..., 100_hist

XGBoost 모델을 최적화하기 위해 대상 n_estimator는 [25, 50, 100], 트리 구성 알고리즘으로는 모든 변인을 열거하는 정확한 알고리즘인 ‘exact’, 그래디언트 히스토그램을 이용한 그리디 알고리즘인 ‘approx’, 보다 빠른 히스토그램 최적화 알고리즘인 ‘hist’를 활용하였다. 학습율은 0.2로 지정하였다. 각 포지션에 따라 각각 진행하였으며 5fold CV로 성능을 측정하였다. F1 score가 높아지도록, 전체 특징에서 하나씩 제거하는 방식으로 특징을 추출하였다. 예측 모델은 zscoring을 사용한 XG부스트 모델 1과 minmax를 사용한 XG부스트 모델 2의 2가지 형태로 나타냈다.

표 48. XGBoost 모델 1 투수 및 포수 Hyper parameter

포지션	criterion	zscoring 변인
투수	50_hist	CG, win, save, hold, IBB, WP, WPA, game, FIP, K-BB%, LOB%, AVG , SLG, OPS, P, WHIP, WHIP+, , IP/G, P/IP, P/PA, CYP
포수	50_exact	WAR, HIT, 2B, SF, BB%, K%, PSN, wRC+

CG: 완투, win: 승, save: 세, hold: 홀드, IBB: 고의 4구, WP: 폭투, WPA: 승리 확륙 기여도, game: 출장, FIP: 수비무관 평균자책점, K-BB%:볼넷 하나당 삼진비율, LOB%: 잔루율, AVG: 타율, SLG: 장타율, OPS: 장타율+출루율, P: 투구수, WHIP: 이닝당 출루허용율, WHIP+: 파크팩터 조정 WHIP, IP/G: 게임당 이닝수, P/IP: 이닝당 투구수, P/PA: 타석당 투구수, CYP: 사이영상 포인트 WAR: 대체선수 대비 승리기여도, HIT: 안타, 2B: 2루타, SF: 희비, BB%: 볼넷율, K%: 삼진율, PSN: 호타준족 점수, wRC+: 파크팩터 조정 wRC

투수의 XGBoost 구현 모델 1 설정값은 50_hist를 조정하여 모형의 일반화 성능을 최적화하였다. 변인 도출 결과, CG, win, save, hold, IBB, WP, WPA, game, FIP, K-BB%, LOB%, AVG , SLG, OPS, P, WHIP, WHIP+, IP/G, P/IP, P/PA, CYP가 변인으로 추출되었다. 포수는 50_exact의 설정값을 조정하여 XGBoost 모델 1 모형을 최적화하였다. 변인 도출 결과, WAR, HIT, 2B, SF, BB%, K%, PSN, WRC+로 나타났다.

표 49. XGBoost 모델 1 내야수 Hyper parameter

포지션	criterion	zscoring 변인
1루수	25_approx	PA, runs, 3B, HBP, SF, wRC+, WPA, BB/K, IsoP, wRC, wRAA
2루수	50_exact	WAR, runs, SO, GIDP, SF, IsoP, wRAA
3루수	25_exact	WAR, TB, HBP, GIDP, SLG, HR%, BB%, PA, wRC
유격수	100_exact	3B, RBI, BB, IBB, SF, AVG, BB/K, Spd, wRC/27, wRC+

PA:타석, runs: 득점, 3B: 3루타, HBP: 몸에 맞는볼, SF: 희비, SO: 삼진, GIDP: 병살, SLG: 장타율, HR%: 홈런율, BB%: 볼넷율, RBI: 타점, BB: 볼넷, IBB: 고의 4구, AVG: 타율, Spd: 스피드 스코어, BB/K: 삼진당 볼넷 수, WAR: 대체선수 대비 승리기여도, wRC: 조정 득점 창출력, wRC+: 파크팩터 조정 wRC, wRC/27: 27아웃당 득점 생산력, IsoP: 순수 장타율, wRAA: 리그평균 대비 득점기여도, WPA: 승리 확률 기여도

1루수의 XGBoost 구현 모델 1 설정값은 25_approx를 조정하여 모형의 일반화 성능을 최적화하였다. zscoring 변인 도출 결과, PA, runs, 3B, HBP, SF, wRC+, WPA, BB/K, IsoP, wRC, wRAA로 나타났다. 2루수는 50_exact의 설정값을 조정하여 XGBoost 모델 1 모형을 최적화하였다. zscoring 변인 도출 결과, WAR, runs, SO, GIDP, SF, IsoP, wRAA로 확인되었다. 3루수는 25_exact의 설정값을 조정하여 XGBoost 모델 1 모형을 최적화하였다. zscoring 변인 도출 결과, WAR, TB, HBP, GIDP, SLG, HR%, BB%, PA, wRC가 변인으로 도출되었다. 유격수는 100_exact의 설정값을 조정하여 XGBoost 모델 1 모형을 최적화하였다. zscoring 변인 도출 결과, 3B, RBI, BB, IBB, SF, AVG, BB/K, Spd, wRC/27, wRC+로 확인되었다.

표 50. XGBoost 모델 1 외야수 및 지명타자 Hyper parameter

포지션	criterion	zscoring값 변인
외야수	50_exact	WAR, runs, 3B, TB, RBI, SB, BB, GIDP, SF, OBP, SLG, OPS, WPA, PA, BB%, K%, BB/K, IsoP, IsoD, BABIP, PSN, wRC, wRC/27, wRAA
지명타자	50_approx	HIT, CS, OPS, K%, BABIP, Spd, wRC

WAR: 대체선수 대비 승리기여도, runs: 득점, 3B: 3루타, TB: 총루타, RBI: 타점, SB: 도루, BB: 볼넷, GIDP: 병살, SF: 희비, OBP: 출루율, SLG: 장타율, OPS: 장타율+출루율, WPA: 승리확률 기여도, PA: 타석, BB%: 볼넷율, K%: 삼진율, BB/K: 삼진당 볼넷 수, IsoP: 순수 장타율, IsoD: 순수 출루율, BABIP: 인플레이 타구 타율, PSN: 호타준족 점수, wRC: 조정 득점 창출력, wRC/27: 27아웃당 득점 생산력, wRAA: 리그평균 대비 득점기여도, HIT: 안타, CS: 도실, Spd: 스피드 스코어

외야수의 XGBoost 구현 모델 1 설정값은 50_exact를 조정하여 모형의 일반화 성능을 최적화하였다. 변인 도출 결과, WAR, runs, 3B, TB, RBI, SB, BB, GIDP, SF, OBP, SLG, OPS, WPA, PA BB%, K%, BB/K, IsoP, IsoD, BABIP, PSN, wRC, wRC/27, wRAA로 나타났다. 지명타자는 50_approx의 설정값을 조정하여 XGBoost 모델 1 모형을 최적화하였다. 변인 도출 결과, HIT, CS, OPS, K%, BABIP, Spd, wRC가 변인으로 확인되었다.

XGBoost 모델 1의 포지션별 중요 변인을 <그림 19>, <그림 20>, <그림 21>과 같이 구현하였다. 투수 중요 변인으로는 승리, 고의4구, 폭투, 게임당 이닝 수 순으로 중요 변인이 확인되었다. 포수는 안타가 42.21%로 가장 중요한 변인으로 추출되었으며 WAR, 희비 순으로 중요 변인이 나타났다. 1루수는 wRAA로 34.35%로 가장 중요한 변인으로 확인되었으며 wRC+, wRC 순으로 높은 중요도로 나타났음을 알 수 있다. 2루수는 WAR이 62.36%로 가장 중요한 변인으로 추출되었으며 3루수는 wRC, 장타율, HR% 순으로 중요 변인이 나타났다. 유격수는 wRC/27이 35.57%로 중요 변인으로 추출되었으며 타율 희비 순으로 확인되었다. 외야수는 WAR, IsoD, 3루타, IsoP 순으로 중요 변인이 확인되었으며, 지명타자는 안타가 47.63%로 가장 중요한 변인으로 추출되었으며, OPS, BABIP, K% 순으로 중요 변인이 나타났다.

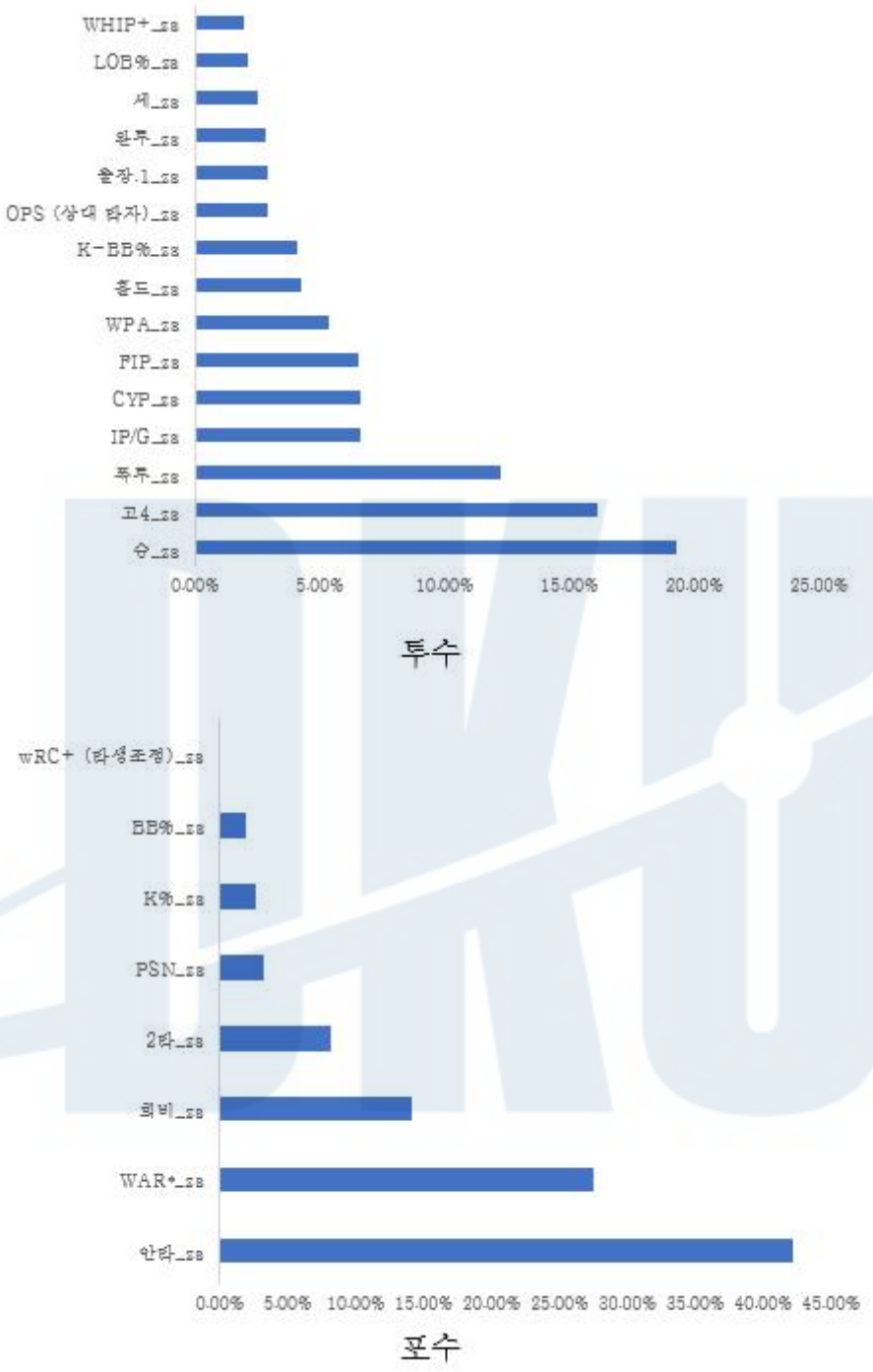


그림 19. XGBoost 모델 1 투수, 포수 변인 중요도

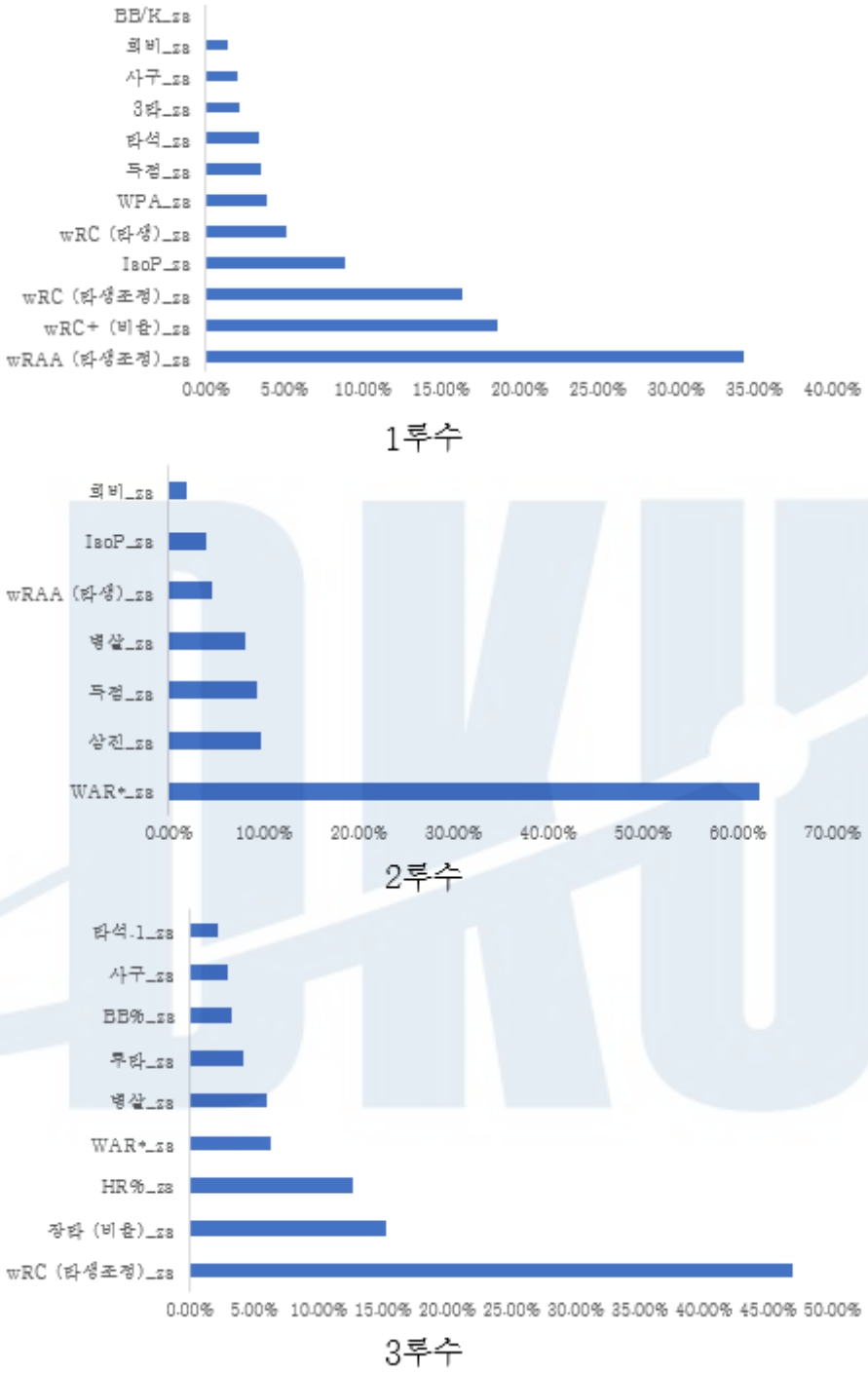


그림 20. XGBoost 모델 1 1루수, 2루수, 3루수 변인 중요도

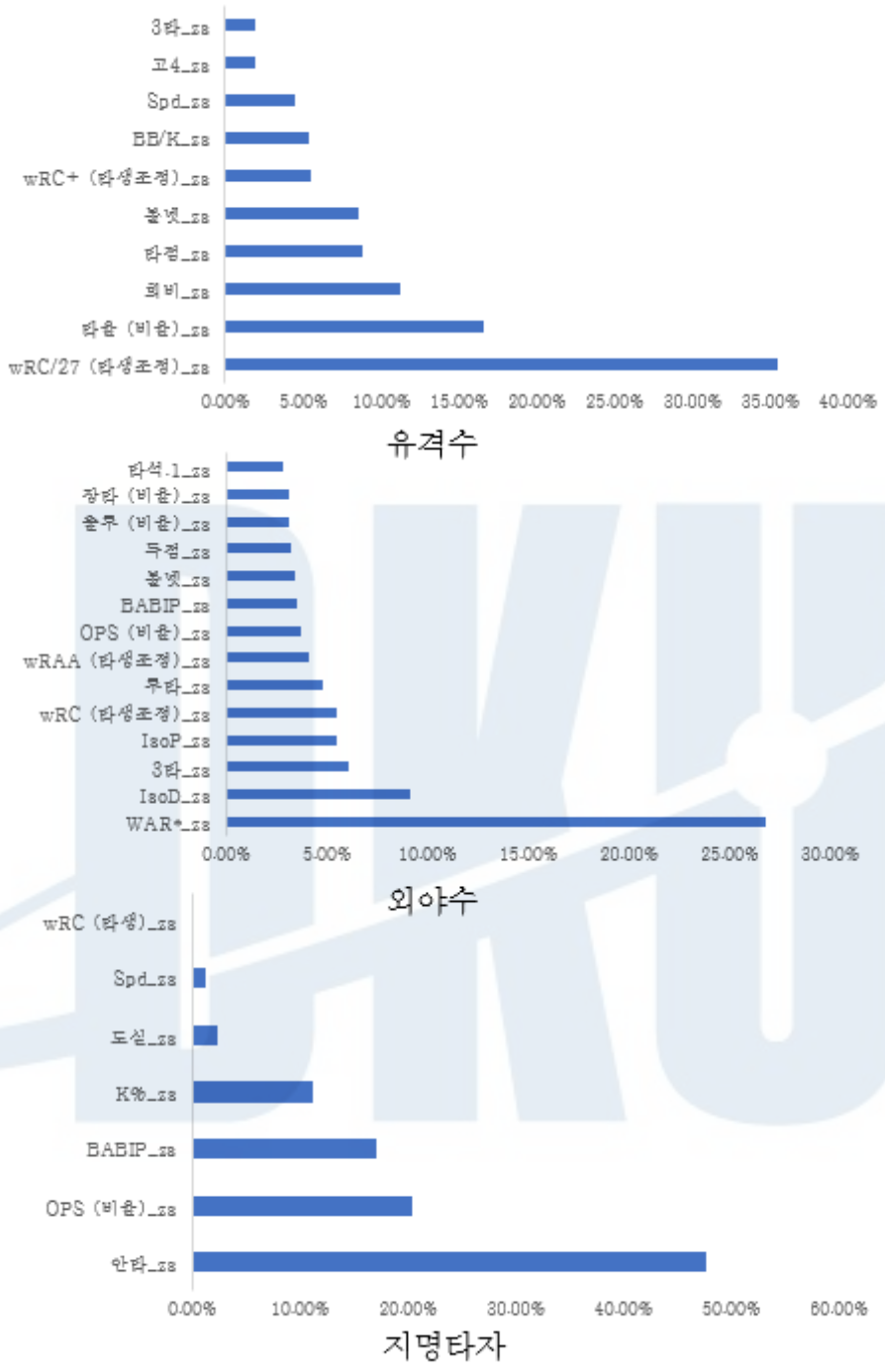


그림 21. XGBoost 모델 1 유격수, 외야수, 지명타자 변인 중요도

표 51. XGBoost 모델 2 투수 및 포수 Hyper parameter

포지션	criterion	minmax 변인
투수	25_approx	win, ERA, PA, FIP, K/BB, LOB%, OPS, P, IP/G, P/IP, CYP
포수	25_exact	SB, IBB, OPS, PA, K%, wOBA, wRC, wRAA

win: 승, ERA: 평균자책점, PA: 타석, FIP: 수비무관 평균자책점, K/BB: 볼넷당 삼진 수, LOB%: 잔루율, OPS: 장타율+출루율, P: 투구수, IP/G: 게임당 이닝 수, P/IP: 이닝당 투구 수, CYP: 사이영상 포인트, SB: 도루, IBB: 몸에 맞는볼, K%: 삼진율, wOBA: 가중출루율, wRC: 조정 득점 창출력, wRAA: 리그평균대비 득점기여도

투수의 XGBoost 구현 모델 2 설정값은 25_approx를 조정하여 모형의 일반화 성능을 최적화하였다. 변인 도출 결과, win, ERA, PA, FIP, K/BB, LOB%, OPS, P, IP/G, P/IP, CYP가 변인으로 추출되었다. 포수는 25_exact의 설정값을 조정하여 XGBoost 모델 2 모형을 최적화하였다. 변인 도출 결과, SB, IBB, OPS, PA, K%, wOBA, wRC, wRAA로 확인되었다.

표 52. XGBoost 모델 2 내야수 Hyper parameter

포지션	criterion	minmax 변인
1루수	25_exact	WPA, PA, BB/K, IsoP, wRAA, wRC
2루수	50_approx	SO, Spd, wRAA, wOBA, wRC/27
3루수	50_approx	WAR, TB, HBP, IBB, SF, IsoP, wRC+
유격수	25_exact	SB, SAC, SF, AVG, WPA, BB%, IsoD, PSN, wOBA, wRC, wRC/27

WPA: 승리 확률 기여도, PA: 타석, BB/K: 삼진당 볼넷 수 IsoP: 순수장타율, wRAA: 리그평균대비 득점기여도, wRC: 조정 득점 창출력, SO: 삼진, Spd: 스피드 스코어, wOBA: 가중 출루율, wRC/27: 27아웃당 득점 생산력, WAR: 대체선수 대비 승리기여도, TB: 총루타, HBP: 몸에 맞는볼, IBB: 고의 4구, SF: 희비, wRC+: 파크팩터 조정 wRC, SB: 도루, SAC: 희타, AVG: 타율, BB%: 볼넷율, IsoD: 순수 출루율, PSN: 호타준족 점수

1루수의 XGBoost 구현 모델 2 설정값은 25_exact를 조정하여 모형의 일반화 성능을 최적화하였다. minmax 변인 도출 결과, WPA, PA, BB/K, IsoP, wRAA, wRC로 확

인되었다. 2루수는 50_approx의 설정값을 조정하여 XGBoost 모델 2 모형을 최적화하였다. minmax 변인 도출 결과, SO, Spd, wRAA, wOBA, wRC/27이 도출되었다. 3루수는 50_approx의 설정값을 조정하여 XGBoost 모델 2 모형을 최적화하였다. minmax 변인 도출 결과, WAR, TB, HBP, IBB, SF, IsoP, wRC+가 나타났음을 알 수 있다. 유격수는 25_exact의 설정값을 조정하여 XGBoost 모델 2 모형을 최적화하였다. minmax 변인 도출 결과, SB, SAC, SF, AVG, WPA, BB%, IsoD, PSN, wOBA, wRC, wRC/27이 나타났다.

표 53. XGBoost 모델 2 외야수 및 지명타자 Hyper parameter

포지션	criterion	minmax값 변인
외야수	50_exact	WAR, Game, runs, 2B, 3B, TB, SB, IBB, GIDP, AVG, wOBA, WPA, PA, IsoP, IsoD, BABIP, PSN, wRC, wRC/27
지명타자	50_approx	PA, runs, HIT, 2B, HR, TB, RBI, HBP, AVG, SLG, OPS, K%, BB/K, IsoD, BABIP, wRC

WAR: 대체선수대비 승리기여도, Game: 출장, runs: 득점, 2B: 2루타, 3B: 3루타, TB: 총루타, SB: 도루, IBB: 고의4구, GIDP: 병살, AVG: 타율, wOBA: 가중 출루율, WPA: 승리 확률 기여도, PA: 타석, IsoP: 순수 장타율, IsoD: 순수 출루율, BABIP: 인플레이 타구 타율, PSN: 호타준족 점수, wRC: 조정 득점 창출력, wRC/27: 27아웃당 득점 생산력, HIT: 안타, HR: 홈런, RBI: 타점, HBP: 몸에 맞는볼, SLG: 장타율, OPS: 장타율+출루율, K%: 삼진율, BB/K: 삼진당 볼넷 수

외야수의 XGBoost 구현 모델 2 설정값은 50_exact를 조정하여 모형의 일반화 성능을 최적화하였다. 변인 도출 결과, WAR, Game, runs, 2B, 3B, TB, SB, IBB, GIDP, AVG, wOBA, WPA, PA, IsoP, IsoD, BABIP, PSN, wRC, wRC/27가 추출되었다.지명타자는 50_approx의 설정값을 조정하여 XGBoost 모델 1 모형을 최적화하였다. 변인 도출 결과 PA, runs, HIT, 2B, HR, TB, RBI, HBP, AVG, SLG, OPS, K%, BB/K, IsoD, BABIP, wRC로 나타났다.

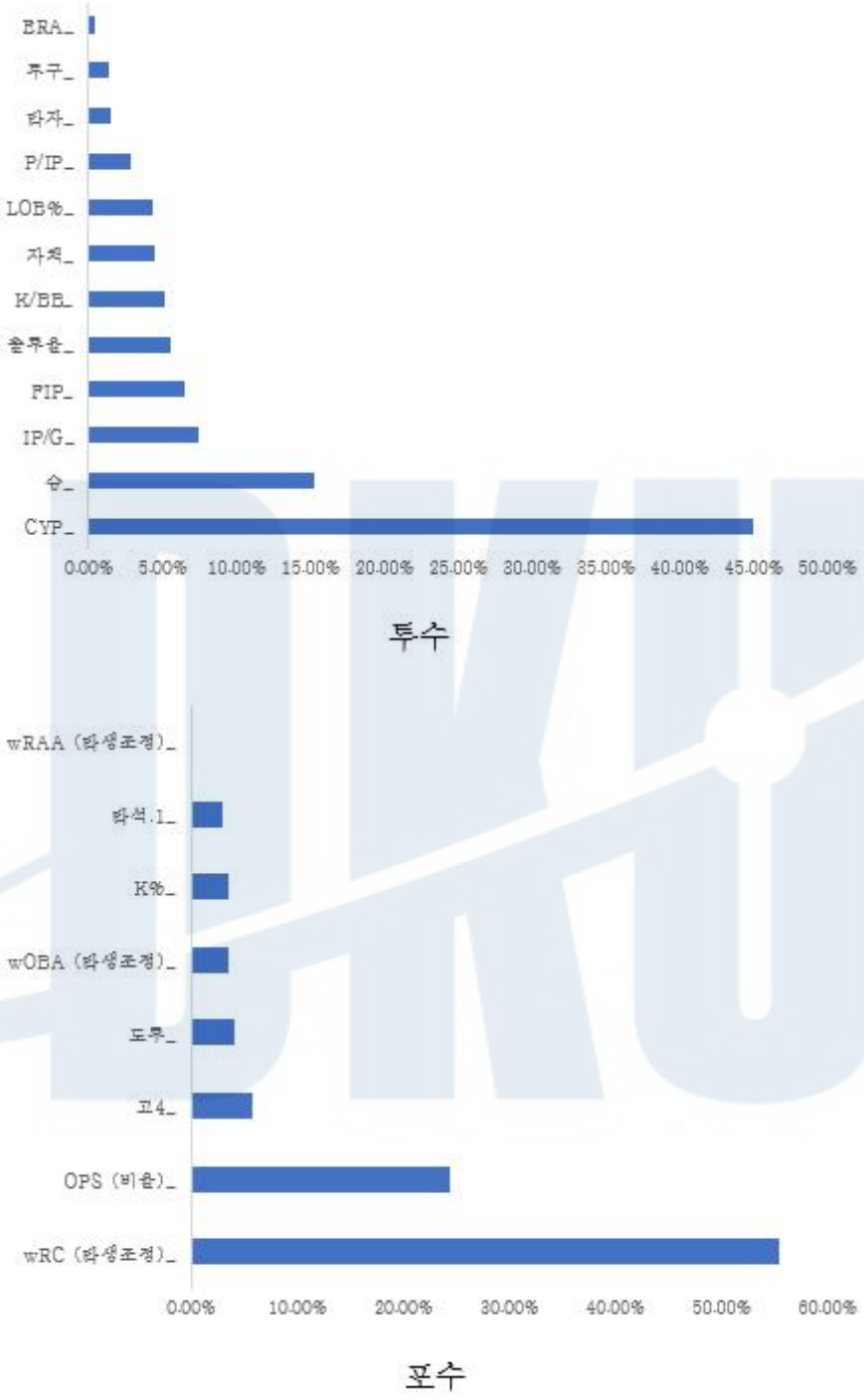


그림 22. XGboost 모델 2 투수, 포수 변인 중요도

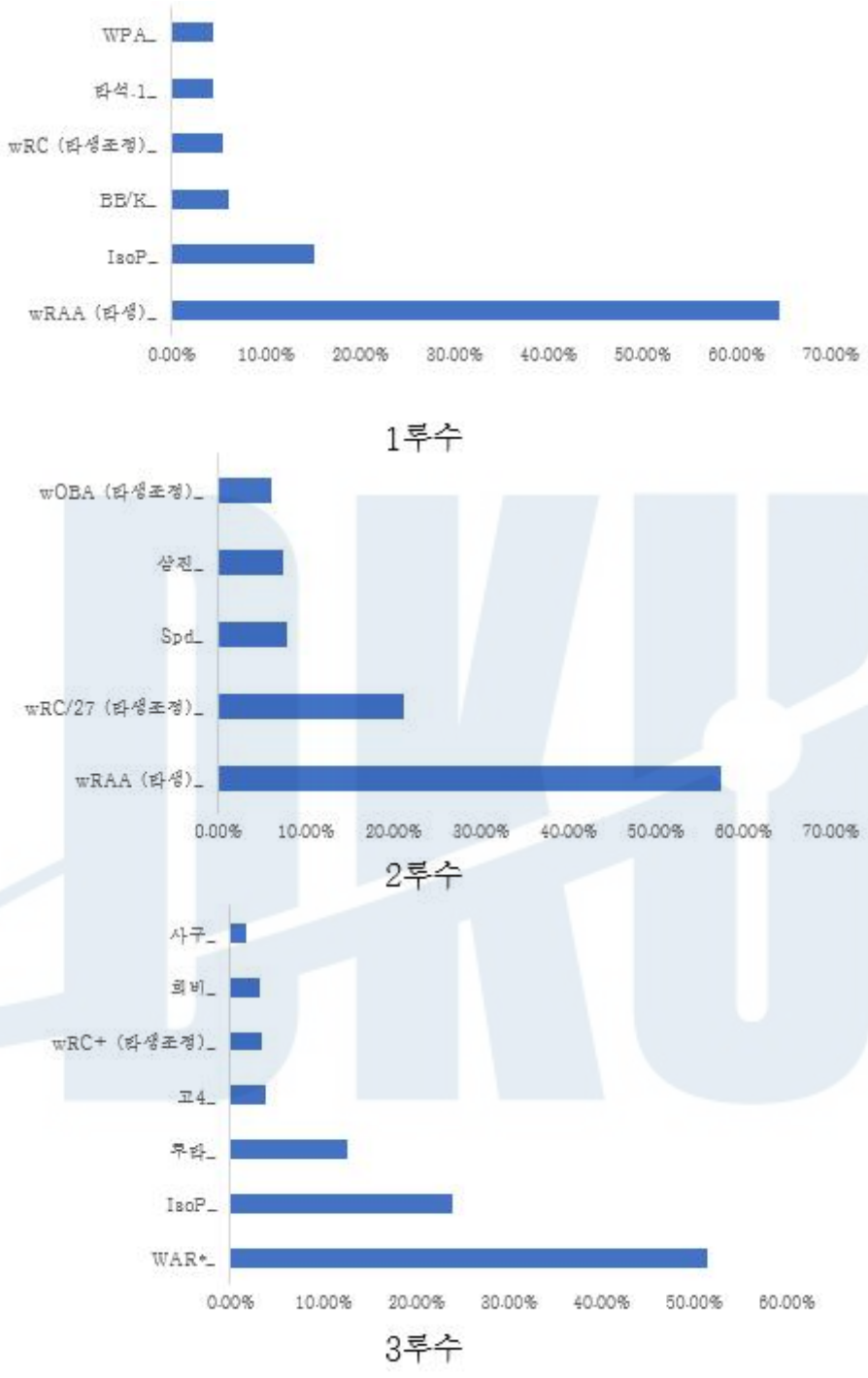


그림 23. XGBoost 모델 2 1루수, 2루수, 3루수 변인 중요도

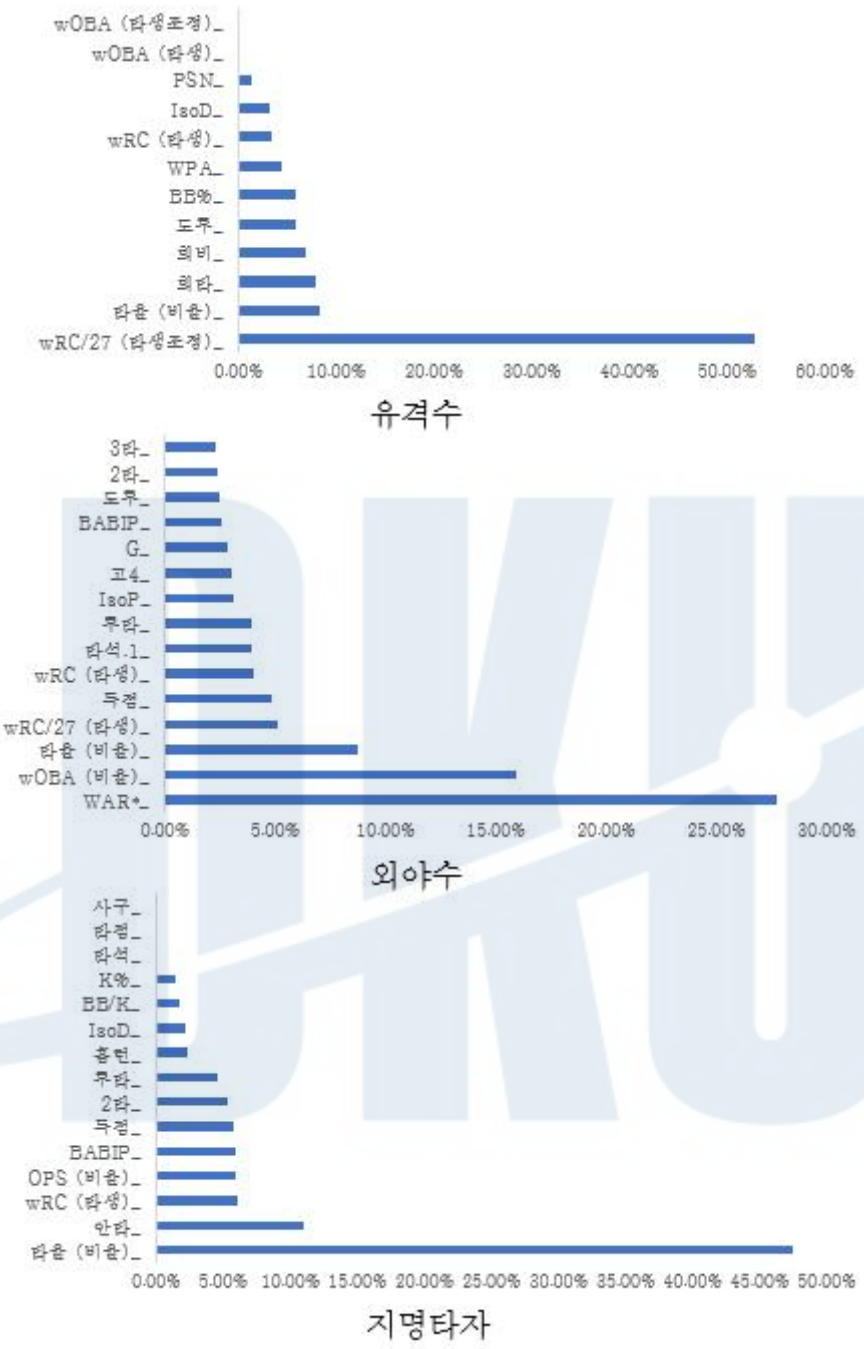


그림 24. XGBoost 모델 2 유격수, 외야수, 지명타자 변인 중요도

XGBoost 모델 2의 포지션별 중요 변인을 <그림 22>, <그림 23>, <그림 24>와 같이 구현하였다. 투수의 중요 변인을 보면 CYP가 44.96%로 가장 중요한 변인으로 추출되었으며, 승리, 게임당 이닝 수, FIP 순으로 중요 변인이 확인되었다. 포수는 wRC가 55.40%로 가장 중요한 변인으로 추출되었으며, OPS가 24.38%로 나타났으며, 고의 4구 순으로 중요 변인이 나타났다. 1루수는 wRAA가 64.56%로 가장 중요한 변인으로 추출되었으며 IsoP 15.12%, BB/K 6.05% 순으로 중요 변인이 확인되었다. 2루수는 5개의 변인 중 wRAA가 57.40%로 가장 중요한 변인으로 알 수 있었으며, wRC/27, Spd, 삼진, wOBA순으로 중요 변인이 추출되었다. 3루수는 WAR이 51.48%로 가장 중요한 변인으로 확인되었으며, IsoP, 루타, 고의 4구 순으로 중요 변인이 추출되었다. 유격수는 wRC/27이 가장 중요한 변인으로 추출되었으며, 타율, 희타, 희비, 도루 순으로 중요 변인이 나타났다. 외야수는 WAR이 27.77%로 가장 중요한 변인으로 나타났으며, wOBA, 타율, wRC/27 순으로 중요 변인이 추출되었다. 지명타자는 타율이 47.58%로 가장 중요한 변인으로 확인되었으며, 안타, wRC, OPS, BABIP 순으로 중요 변인이 추출되었다.

2) 로지스틱 회귀분석 및 머신러닝 예측 모델별 성능평가 결과

(1) 로지스틱 회귀분석 모델 포지션별 성능평가 결과

로지스틱 회귀분석의 포지션별 성능평가는 zscoring을 사용한 로지스틱 회귀분석 모델1과 minmax를 사용한 로지스틱 회귀분석 모델2로 구분하여 실시하였다. 정확도(accuracy), 민감도(sensitivity), 특이도(specificity), 재현율(recall), 정밀도(precision), F1 score 지표를 사용하여 성능평가를 실시하였다.

표 54. 각 포지션별 로지스틱 회귀분석 모델별 성능평가 결과표

구분	로지스틱 회귀분석 모델1						로지스틱 회귀분석 모델2					
	acc	sen	spe	rec	pre	F1	acc	sen	spe	rec	pre	F1
투수	0.500	0.550	0.988	0.550	0.767	0.638	0.700	0.700	1.000	0.700	1.000	0.810
포수	0.820	0.850	0.986	0.850	0.950	0.893	0.820	0.850	0.986	0.850	0.950	0.893
1루수	0.860	0.900	0.987	0.900	0.960	0.921	0.910	0.950	0.988	0.950	0.960	0.949
2루수	0.810	0.850	0.987	0.850	0.960	0.883	0.820	0.850	0.987	0.850	0.950	0.893
3루수	0.760	0.850	0.962	0.850	0.843	0.842	0.723	0.850	0.949	0.850	0.827	0.824
유격수	0.730	0.800	0.973	0.800	0.910	0.842	0.750	0.850	0.948	0.850	0.840	0.838
외야수	0.652	0.721	0.975	0.721	0.894	0.785	0.645	0.735	0.966	0.735	0.875	0.780
지명타자	0.860	0.900	0.985	0.900	0.960	0.921	0.820	0.900	0.969	0.900	0.920	0.889

acc: 정확도, sen: 민감도, spe: 특이도, rec: 재현율, pre: 정밀도

로지스틱 회귀분석 모델 1의 성능평가를 결과를 보면 투수와 외야수가 0.500, 0.652로 가장 낮은 정확도로 확인되었으며, 1루수, 지명타자가 0.860으로 가장 높은 정확도로 확인되었다. 정밀도는 1루수, 2루수, 3루수가 0.960으로 가장 높게 나타났으며, 투수가 0.767로 가장 낮게 나타났다. F1 score는 1루수, 지명타자가 0.921로 높은 경향을 보였으며, 포수, 2루수, 3루수, 유격수가 0.8대, 투수가 0.638로 확인되었다. 로지스틱 회귀분석 모델 2는 1루수가 0.910으로 가장 높은 정확도로 확인되었으며, 외야수가 0.645로 확인되었다. 재현율은 1루수, 지명타자가 0.9대, 포수, 2루수, 3루수, 유격수가 0.850, 외야수, 투수가 0.735, 0.700으로 확인되었다. F1 score는 1루수, 2루수, 포수, 지명타자 순으로 확인되었다. 로지스틱 회귀분석 모델 2가 모델 1보다 투수, 1루수, 2루수, 유격수에서 높은 정확도를 보였으며, 3루수, 외야수, 지명타자는 낮은 정확도로 나타났음을 알 수 있다.

(2) 서포트 백터 머신 모델 포지션별 성능평가 결과

서포트 백터 머신의 포지션별 성능평가는 zscoring을 사용한 서포트 백터 머신 모델1과 minmax를 사용한 서포트 백터 머신2로 구분하여 실시하였다. 정확도(accuracy), 민감도(sensitivity), 특이도(specificity), 재현율(recall), 정밀도(precision), F1 score 지표를 사용하여 성능평가를 실시하였다.

표 55. 각 포지션별 서포트 백터 머신 모델별 성능평가 결과표

구분	서포트 백터 머신 모델1						서포트 백터 머신 모델2					
	acc	sen	spe	rec	pre	F1	acc	sen	spe	rec	pre	F1
투수	0.350	0.350	0.996	0.350	0.800	0.480	0.810	0.850	0.996	0.850	0.960	0.883
포수	0.800	0.800	1.000	0.800	1.000	0.886	0.920	0.950	0.986	0.950	0.950	0.950
1루수	0.900	0.900	1.000	0.900	1.000	0.943	0.910	0.950	0.988	0.950	0.960	0.949
2루수	0.720	0.750	0.987	0.750	0.950	0.826	0.800	0.800	1.000	0.800	1.000	0.876
3루수	0.730	0.800	0.974	0.800	0.910	0.833	0.870	0.900	0.987	0.900	0.950	0.921
유격수	0.720	0.750	0.987	0.750	0.950	0.826	0.750	0.800	0.973	0.800	0.920	0.848
외야수	0.602	0.687	0.966	0.687	0.861	0.744	0.663	0.769	0.957	0.769	0.832	0.793
지명타자	0.780	0.850	0.969	0.850	0.910	0.871	0.820	0.900	0.969	0.900	0.920	0.898

acc: 정확도, sen: 민감도, spe: 특이도, rec: 재현율, pre: 정밀도

서포트 백터 머신 모델 1의 성능평가를 결과를 보면 투수가 정확도 0.350으로 가장 낮게 나타났으며, 1루수가 정확도 0.900, F1 score 0.943으로 가장 높게 확인되었다. 2루수, 3루수, 유격수, 지명타자 포지션 모두 0.7대의 정확도로 확인되었으며, 0.8 이상의 F1 score로 나타났다. 재현율은 1루수, 지명타자, 3루수, 2루수 순으로 확인되었다. 서포트 백터 머신 모델 2는 포수가 0.920로 가장 높은 정확도를 보였으며, 외야수가 0.663으로 가장 낮은 정확도를 보였다. 정밀도는 2루수 1.000으로 가장 높게 나타났으며, 외야수를 제외한 모든 포지션에서 0.9이상으로 확인되었다. F1 score는 포수, 1루수, 3루수가 0.9 이상으로 높은 경향을 보였으며, 지명타자, 투수, 2루수, 유격수가 0.8이상, 외야수가 0.793으로 나타났다. 전체적으로 서포트 백터 머신 모델 1보다 서포트 백터 머신 모델 2가 더 정확한 예측모델임을 알 수 있다.

(3) 랜덤포레스트 모델 포지션별 성능평가 결과

랜덤포레스트의 포지션별 성능평가는 zscoring을 사용한 랜덤포레스트 모델1과 minmax를 사용한 랜덤포레스트 모델2로 구분하여 실시하였다. 정확도(accuracy), 민감도(sensitivity), 특이도(specificity), 재현율(recall), 정밀도(precision), F1 score 지표를 사용하여 성능평가를 실시하였다.

표 56. 각 포지션별 랜덤포레스트 모델별 성능평가 결과표

구분	랜덤포레스트 모델1						랜덤포레스트 모델2					
	acc	sen	spe	rec	pre	F1	acc	sen	spe	rec	pre	F1
투수	0.440	0.450	0.996	0.450	0.900	0.585	0.590	0.650	0.992	0.650	0.893	0.715
포수	0.750	0.850	0.957	0.850	0.860	0.849	0.780	0.850	0.971	0.850	0.910	0.861
1루수	0.753	0.850	0.962	0.850	0.883	0.853	0.820	0.900	0.974	0.900	0.920	0.898
2루수	0.693	0.800	0.960	0.800	0.893	0.814	0.720	0.800	0.973	0.800	0.920	0.832
3루수	0.750	0.850	0.962	0.850	0.860	0.849	0.790	0.850	0.974	0.850	0.893	0.864
유격수	0.550	0.650	0.948	0.650	0.803	0.702	0.520	0.700	0.909	0.700	0.710	0.683
외야수	0.594	0.686	0.966	0.686	0.871	0.731	0.608	0.687	0.957	0.687	0.815	0.740
지명타자	0.720	0.750	0.985	0.750	0.950	0.826	0.690	0.750	0.969	0.750	0.893	0.806

acc: 정확도, sen: 민감도, spe: 특이도, rec: 재현율, pre: 정밀도

랜덤포레스트 모델 1의 성능평가를 결과를 보면 1루수, 3루수, 포수가 0.75대로 가장 높은 정확도를 보였으며, 투수, 유격수 외야수가 0.440, 0,550. 0.594로 낮은 정확도로 나타났다. 재현율은 포수, 1루수, 3루수가 높게 나타났으며, 투수가 가장 낮게 나타났다. 정밀도는 지명타자, 투수, 2루수 순으로 확인되었으며, 유격수가 가장 낮게 나타났다. F1 score는 1루수가 0.853으로 가장 높게 나타났으며 투수가 0.585로 가장 낮은 수치를 보였다. 랜덤포레스트 모델 2는 1루수가 가장 높은 정확도를 보였으며, 유격수가 가장 낮은 정확도를 보였다. 정밀도는 1루수, 2루수가 0.920으로 가장 높게 나타났으며, 유격수가 0.710으로 가장 낮게 나타났다. F1 score는 1루수, 3루수, 포수 순으로 높게 확인되었으며 유격수가 가장 낮게 나타났음을 알 수 있다. 랜덤포레스트 모델 2가 모델1에 비해 6개 포지션(투수, 포수, 1루수, 2루수, 3루수, 외야수)에서 정확도가 높게 나타났으며, 2개 포지션(유격수, 지명타자)에서 정확도가 낮은 경향을 보였다.

(4) XGBoost 모델 포지션별 성능평가 결과

XGBoost 모델의 포지션별 성능평가는 zscoring을 사용한 XGBoost 모델1과 minmax를 사용한 XGBoost 모델2로 구분하여 실시하였다. 정확도(accuracy), 민감도(sensitivity), 특이도(specificity), 재현율(recall), 정밀도(precision), F1 score 지표를 사용하여 성능평가를 실시하였다.

표 57. 각 포지션별 XGBoost 모델별 성능평가 결과표

구분	XGBoost 모델1						XGBoost 모델2					
	acc	sen	spe	rec	pre	F1	acc	sen	spe	rec	pre	F1
투수	0.620	0.650	0.996	0.650	0.950	0.760	0.767	0.800	0.992	0.800	0.900	0.833
포수	0.820	0.900	0.971	0.900	0.920	0.898	0.870	0.950	0.970	0.950	0.920	0.927
1루수	0.860	0.900	0.987	0.900	0.960	0.921	0.870	0.950	0.974	0.950	0.920	0.927
2루수	0.760	0.800	0.987	0.800	0.960	0.844	0.760	0.800	0.987	0.800	0.960	0.864
3루수	0.870	0.900	0.988	0.900	0.950	0.921	0.780	0.850	0.973	0.850	0.910	0.861
유격수	0.614	0.700	0.960	0.700	0.914	0.755	0.700	0.800	0.948	0.800	0.840	0.800
외야수	0.646	0.722	0.970	0.722	0.874	0.782	0.693	0.804	0.957	0.804	0.841	0.818
지명타자	0.880	0.950	0.969	0.950	0.910	0.928	0.840	0.900	0.969	0.900	0.893	0.892

acc: 정확도, sen: 민감도, spe: 특이도, rec: 재현율, pre: 정밀도

XGBoost 모델 1의 성능평가 결과를 보면 지명타자, 3루수, 1루수 순으로 높은 정확도를 보였으며, 유격수, 투수, 외야수가 낮은 정확도를 보였다. 정밀도는 1루수, 2루수, 3루수, 투수가 0.95이상으로 나타났으며, 외야수가 0.874로 가장 낮게 나타났다. F1 score는 지명타자가 가장 0.928로 가장 높음을 확인할 수 있었고, 1루수, 3루수, 포수 순으로 확인되었다. 유격수, 투수, 외야수가 0.7대로 낮게 나타났다. XGBoost 모델 2에서는 포수, 1루수, 지명타자 순으로 정확도가 높게 나타났으며, 외야수가 0.693으로 가장 낮은 정확도로 확인되었다. 재현율은 포수, 1루수가 가장 높게 나타났으며, 투수, 2루수, 유격수가 0.800으로 낮게 나타났다. F1 score는 포수, 1루수가 0.927로 가장 높게 나타났으며, 유격수가 0.800으로 낮게 나타났다. XGBoost 모델 1에 비해 모델 2가 투수, 포수, 1루수, 유격수, 외야수에서 정확도가 높게 나타났으며, 3루수, 지명타자는 낮게 나타났음을 알 수 있다.

Ⅴ. 논의

이 연구는 2003년~2022년까지 한국프로야구 골든글러브 후보 선수 및 수상자 선수의 기록을 기반으로 로지스틱 회귀분석과 머신러닝을 활용한 골든글러브 수상 예측모델을 설계하고, 설계된 예측모델의 성능을 비교, 분석하여 골든글러브 수상에 적합한 모델을 찾는데 목적이 있다. 이 연구의 목적을 달성하기 위해 로지스틱 회귀분석과 머신러닝 예측기법인 서포트 백터 머신, 랜덤포레스트. XGBoost 모델을 설계하고 모델별로 최적화 변인 탐색 후 성능평가를 실시하였다. 이에 선행연구를 바탕으로 연구결과의 의미를 각 포지션별 골든글러브 후보 및 수상자 간 기록 비교, 각 포지션별 최적화 변인탐색, 예측모델의 성능 비교 및 분석으로 다음과 같이 논의하고자 한다.

1. 골든글러브 후보 및 수상자 간 포지션별 기록 비교

골든글러브 수상자 예측 모델을 개발하기 전 각 포지션별 전체 선수 기록, 수상그룹 기록, 미수상 그룹 기록으로 구분하여 기술통계를 실시하였다. 야구에서 세이버 메트릭스 지표가 국내에 도입되기 전 투수의 능력을 가장 높게 평가하는 지표로 승리, 방어율, 탈삼진 항목을 가장 높게 평가하였다. 본 연구에서도 승리 수는 전체 평균 10.1승, 미수상 그룹 9.6승에 비해 수상 그룹 17승으로 높게 나타났으며, 방어율 부분에서도 전체 그룹 및 미수상 그룹 3.41에 비해 수상 그룹 2.72로 낮음을 알 수 있었으며, 탈삼진 수는 전체 평균 116.3개, 미수상 그룹 평균 112.9개, 수상 그룹 평균 158.2개로 확인되었다. 이장택(2014)의 ‘한국프로야구에서의 투수평가지표’ 연구에서는 ERA, K/9, BB/9, WHIP, BABIP, FIP 등의 기록을 중요시하여 투수지표를 개발하였다. WHIP와 K/9, BB/9, WHIP, FIP에서 수상 그룹이 뛰어나게 나타났으나, BABIP는 인플레이 타구의 타율로 운이 따르는 지표로서, 전체평균, 미수상 그룹, 수상그룹이 비슷하게 확인되었다. 타자 부분을 보면 양도업, 조은형, 배상우, 정상원

(2015)의 연구에서 타자의 경기력요인을 분석하기 위하여 OPS, RC, ISOP, PSN, wOBA, BABIP, TA, RC등의 지표를 사용하여 요인점수로 나타냈다. 이장택(2014)의 '한국프로야구에서 타자능력의 측정' 에서는 TB, OBP, SLG, OPS, IsoP, wOBA, XR 등의 기록으로 타자지표를 제안하였다. 1루수는 타격능력이 뛰어나고 장타력이 뛰어난 타자들이 많은 포지션이다. 1루수는 홈런왕이 많이 탄생한 포지션으로 홈런, 장타율, IsoP가 골든글러브 수상 그룹이 전체, 미수상 그룹, 그리고 타 포지션에 비해 가장 높게 나타났음을 알 수 있다. 2루수는 정교한 타격과 작전 수행능력이 뛰어난 포지션으로서, 타율부분에서 전체 그룹 및 미수상 그룹보다 뛰어남을 알 수 있었으며, 희생타 수에서는 타 포지션에 비해 작전수행이 많은 포지션으로서, 희생타 수가 높게 나타났고, 미수상 그룹 및 전체 그룹에 비해 수상 그룹의 희생타 수가 적게 나타났는데 이는 골든글러브 수상선수의 타격능력 즉, 컨택이 뛰어나기 때문에 희생타 수가 적었음을 알 수 있다. 3루수는 한국프로야구에서 1루수 다음으로 홈런타자가 많이 배출된 포지션이다. 홈런, 장타율, IsoP, OPS가 높게 나타났으며, 장타가 많다보니 총 루타수도 높게 나타났다. 유격수는 한국프로야구에서 수비능력이 뛰어난 포지션으로 전체적으로 타격능력은 저조하나 전체 및 수상 그룹에서 좋은 타격능력을 보였다. 이는 현재 메이저리그에서 뛰고 있는 김하성 선수나 예전에 뛰었던 강정호 선수가 한국프로야구에서 뛰어난 타격성적을 거두었기에, 전체적으로 뛰어난 타격성적이 기록되었다. 가장 많은 선수가 수상하는 외야수는 전체적인 타격기록에 대한 밸런스가 뛰어나거나 타이틀홀더(홈런, 득점, 최다안타 등)가 수상자가 되어왔다. 수상 그룹이 안타 수 160.9개로 타 포지션에 비해 가장 많은 수치로 나타났으며, 득점, 도루 등에서 타 포지션보다 뛰어난 것으로 확인되었다. 포수 포지션에서는 2011년부터 2명의 선수가 골든글러브를 계속 수상해왔다. 2명의 선수의 타격 성적은 타 포지션과 비슷했으나 전체적인 포수의 타격성적은 낮음을 알 수 있다. 이는 야구에서 체력소모가 가장 심한 포수 포지션이 여름철 더위, 길어지는 경기 시간 등으로 인해 타격성적이 감소했음을 알 수 있다. 지명타자는 타격능력이 뛰어난 선수들이 출전하던 포지션에서 최근에는 고참 선수나 체력안배를 위한 포지션으로 많이 사용되고 있다. 타격능력이 뛰어난 선수가 많이 출전하면서 타율은 전

포지션 중 가장 뛰어나게 나타났다. 한국프로야구에서는 골든글러브를 수상하는데 수비능력보다는 타격능력을 극대화하여 수상을 진행해 왔다. 2023년부터 ‘수비상’이 신설되었다. 허주한, 우용태(2023)의 ‘한국프로야구에서 빅 데이터 분석 기법을 이용한 수비수 평가모델’의 연구에서 수비수의 수비능력을 객관화하기 위한 새로운 모델을 제시하였는데 한국프로야구에서도 미국 메이저리그와 새로운 수비 평가모델 개발 자료를 토대로 ‘수비상’이 아닌 골든글러브 수상으로의 변화를 기대한다. 또한 골든글러브 수상이 기자단투표를 통해 수상자가 결정되다보니 인기투표로 변질되는 경우가 있는데 이에 대한 대책 마련도 필요하다고 사료된다.

2. 로지스틱 회귀분석 및 머신러닝 예측모델 분석 결과

1) 로지스틱 회귀분석 및 머신러닝 예측 모델별 최적 변수 결정

로지스틱 회귀분석과 머신러닝(서포트 백터 머신, 랜덤포레스트, XGBoost) 모델별 변인을 결정하였다. zscoring과 minmax를 사용하여 표준화한 모델1과 모델2로 각각 나타냈다. 각 포지션별로 F1 score가 높아지도록 하나씩 빼가면서(Backward)로 특징을 추출한 결과 다양한 변인들이 존재하였다. 투수 변인으로는 승리 수, 패배 수, 폭투, 게임 수, 평균자책점, 상대타자 수, K/BB, 투구 수, 탈삼진 수, 잔루율, 피안타 수, 볼넷 수, 몸에 맞는 볼 수, 피 홈런, 실점 수, 자책점 수 등의 기본 기록의 변인과 FIP, WHIP, WHIP+등의 세이버메트릭스 변인들이 나타났다. 타자 변인으로는 안타 수, 2루타 수, 희생타 수, 삼진 수, 볼넷 수, 병살타 수, 홈런 수, 총 루타수, 장타율등의 기본 기록의 변인과 IsoP, wRAA, WAR, wRC/27, wRC, wRC+, BABIP, WPA, IsoD등의 세이버 메트릭스 변인이 나타났다. 로지스틱 회귀분석 모델에서는 L1, L2, elasticnet이 kernel로 사용되었고, 서포트 백터 머신 모델에서는 rbf, poly가 kernel로 사용되었다. 랜덤포레스트 모델에서는 gini, entropy가 criterion으로 사용되었으며, XGBoost 모델에서는 exact, approx, hist가 criterion으로 사용되었다. 이처럼 본 연구에서는 각 모델별 하이퍼파라미터와 투입 변인들을 상세히 설명하였다. Ji, Zhanga, Shangb & Liu(2021)은 합성곱 신경망 기반의 인코더와 디코더 네트워크 연구에서 연구 설계의 확인을 위해 측정변수 조정을 통한 비교의 필요성을 주장하였고, 김주학, 조선미, 강지연(2022)의 ‘야구 경기 승패 예측을 위한 합성곱 신경망(CNN) 최적화’ 연구에서 야구 경기 승패 예측에 영향을 미치는 요인을 측정 변수 이외에 다른 상황 및 조건이 있어야 한다고 주장하였는데 본 연구에서 하이퍼 파라미터 및 설정 변인을 상세히 제시한 점에서 위의 논문과 같이 측정 변인 및 모델을 조정하여 예측모델을 개발하는 후속연구를 진행 하는데 도움이 될 것이라 사료된다.

2) 로지스틱 회귀분석 및 머신러닝 예측 모델별 성능평가 결과

4가지 머신러닝(로지스틱 회귀분석, 서포트 백터 머신, 랜덤포레스트, XGBoost) 기법의 모델을 각각 zscoring값으로 표준화 시킨 모델1, minmax값으로 표준화 시킨 모델2로 총 8개의 예측모델이 개발되었다. 각각의 모델별 정확도, 민감도, 특이도, 정밀도, 재현율, F1 score를 살펴본 결과 로지스틱 회귀분석 모델2, 서포트 백터 머신 모델 2, 랜덤포레스트 모델2가 정확율과 F1 score가 높게 나타났다. XGBoost 모델의 경우 모델1과 모델 2의 정확도와 F1 score가 비슷하게 나타났다. 투수의 정확도는 서포트 백터 머신 모델 2에서 정확도 0.810, F1 score 0.883으로 가장 높음을 확인할 수 있었다. 가장 많은 골든글러브 후보와 수상자가 있는 외야수의 경우 정확도가 타 포지션에 비해 낮게 나타났는데 XGBoost 모델2가 0.693으로 가장 높게 나타났으며, F1 score 또한 0.818로 가장 높게 나타났음을 알 수 있다. 전체적으로 정확도와 정밀도, F1 score가 높게 나타난 모델은 서포트 백터 머신 모델 2와 XGBoost 모델 2로 확인되었다.

한정섭 등(2022)의 머신러닝을 활용하여 KBO타자의 OPS 예측을 실시한 연구에서 XGBoost, LightGBM, Randomforest, SVR, Linear Regression, Ridge, Lasso와 같은 예측기법을 비교하여 모델의 성능을 평가한 결과 XGBoost가 가장 높게 났으며, 최형준(2022)의 “축구 경기 결과 예측을 위한 머신러닝 기법 비교”에서 로지스틱 회귀분석, 선형판별분석, 인공신경망 모델, 딥러닝 모델, 서포트 백터 머신, 나이브 베이즈 모델, XGBoost 모델을 비교하여 성능을 평가한 결과 선형 판별 분석이 가장 뛰어난 예측성능을 보였고, 서포트 백터머신, XGBoost 모델이 높은 F1 Score가 나타나 예측력이 좋은 것으로 나타났다. 예원진, 이성노(2022)의 FIBA 남자농구 아시안컵 경기결과를 활용한 머신러닝 분류 모형의 예측성능 비교 연구에서 KNN(K Nearest Neighbor), Decision Tree, Support Vector Machine(SVM), LogisticRegression, Random Forest 모델을 비교하여 성능 평가를 실시한 결과 SVM모델이 최적의 예측성능으로 나타났다고 하였는데, 이러한 결과는 서포트 백터머신과 XGBoost 모델에서 높은 예측력을 보인 본 연구결과를 뒷받침해준다. 이처럼 다양한 스포츠 종목에

서 머신러닝 예측기법을 이용한 다양한 연구들이 이루어지고 있는데 연구결과에는 차이가 존재한다. 양민정(2022)은 분석자료의(data)의 양, 특성 그리고 하이퍼 파라미터의 설정에 따라 분석결과가 달라질 수 있다고 하였고, 최형준(2022), 예원진, 이성노(2022), 김주학, 조선미, 강지연(2022)은 동일한 알고리즘 내에서 하이퍼 파라미터를 조절하여 최적의 모델을 개발하였다. 이는 본 연구에서 골든글러브 후보 예측모델의 성능이 높은 수준을 보이지만 데이터의 양이 많아지고 하이퍼 파라미터의 설정을 다르게 한다면 더 뛰어난 예측모델이 개발될 것이라 사료된다.

Ⅵ. 결론 및 제언

1. 결론

이 연구는 2003년~2022년까지 한국프로야구 골든글러브 후보 선수 및 수상자 선수의 기록을 기반으로 머신러닝을 활용한 골든글러브 수상 예측모델을 설계하고, 설계된 예측모델의 성능을 비교, 분석하여 골든글러브 수상에 적합한 모델을 찾는데 목적이 있다. 이 연구의 목적을 달성하기 위해 한국프로야구 2003년부터 2022년까지 골든글러브 후보 및 수상자 선수들의 기록의 정량적 범위를 나타냈고, 로지스틱 회귀분석과 머신러닝 예측기법인 서포트 백터 머신, 랜덤포레스트. XGBoost 모델을 설계하고 모델별로 최적화 변인 탐색 후 성능평가를 실시하였다.

첫째, 한국프로야구 골든글러브 예측모델을 설계하는데 있어 로지스틱 회귀분석 모델에서는 L1, L2, elasticnet이 커넬(kernel)로 사용되었고, 서포트 백터 머신 모델에서는 rbf, poly가 커넬(kernel)로 사용되었으며, 비선형 모델로서 중요 변인은 탐색하지 못하였다. 랜덤포레스트 모델에서는 gini, entropy가 준거(criterion)로 사용되었으며, XGBoost 모델에서는 exact, approx, hist가 criterion으로 사용되었다. F1 score가 높아지도록 하기 위하여 변인을 하나씩 제거하는 방식으로 진행하였고, 각 모델에서 포지션별 사용 변인은 모두 다르게 선정되었다.

둘째, 머신러닝 예측모델의 예측 성능을 비교한 결과 각 포지션별 차이는 존재하지만 서포트 백터 머신 모델 2와 XGBoost 모델의 예측 정확도와 F1 score가 높게 나타났으며, 로지스틱 회귀분석 모델과 랜덤포레스트 모델의 정확도와 F1 score는 상대적으로 낮게 나타났다. 전체적으로 zsocring으로 표준화 한 모델 1보다 minmax로 표준화 한 모델 2의 예측 능력이 뛰어나게 나타났다.

결론적으로 골든글러브 후보 및 수상자의 기록을 기반으로 골든글러브 수상자 예측이 가능하였으며, 각 모델별 중요 변인을 탐색할 수 있었으며, 성능평가를 통해

XGBoost 모델 2가 가장 적합한 모델로 나타났다.

이 연구는 다양한 머신러닝 예측기법을 골든글러브 수상자 예측에 적용했다는 점에서 의미하는 바가 크다. 하지만 머신러닝이 학습을 기반으로 분석이 이루어지기 때문에 분석 자료의 종류와 양에 따라 적용이 어려운 점도 존재하며, 머신러닝을 활용하기 위해서는 하이퍼 파라미터를 설정해야 하는데 하이퍼 파라미터는 연구자, 연구방법, 연구내용에 따라 계속 달라지기 때문에 최적의 하이퍼 파라미터를 설정하기에는 어려움이 존재하며, 하이퍼 파라미터의 설정에 따라 결과가 달라질 수 있다는 문제점을 가지고 있다.

골든글러브 수상은 기자단 투표로 이루어지다보니 기록뿐만 아니라 소속팀의 인기, 선수의 스타성이 골든글러브 수상자 선정에 영향을 미치기 때문에 본 연구에서 나타난 각 모델의 포지션별 중요 변인이 골든글러브 수상에 가장 중요한 변인이라고 결정하는 데는 한계가 존재한다.

2. 제언

추후 연구에서는 다양한 머신러닝 기법으로 모델마다 하이퍼 파라미터를 최적화하는 연구가 진행되어야 한다. 하이퍼 파라미터의 최적화 연구는 준거기준에 따라 상이하게 달라지는데, 최근 등장한 머신러닝 모델인 AutoML(Automated Machine Learning)은 어떤 하이퍼 파라미터가 최적인지를 자동화 해주는 기능을 가지고 있어 추후 연구를 진행하는데 많은 도움이 될 것이라 사료된다. 또한, 본 연구에서는 수비기록을 제외하고 공격기록으로 골든글러브 수상자 예측 연구를 진행하였는데 추후에는 올해부터 생긴 한국프로야구 '수비상'과 같이 수비기록이 동반된 한국프로야구 골든글러브 수상자 예측이 이루어지길 기대한다.

참고문헌

강준만(2009). 한국현대사 산책. 서울: 인물과 사상사.

권순규, 이규원, 최형준(2019). 2016~2018 한국프로야구 세이버 메트릭스 지표 분석, *한국체육과학회지, 28*(3), 1015-1023.

김욱기(2011). 프로야구 구단-소비자의 사회공헌활동 적합성과 팀 동일시, 관계품질(신뢰, 몰입), 팀 충성도와의 구조적 관계. *한국체육학회지, 50*(5), 236-250.

김종훈, 김경태, 한종기(2015). Deep Learning 기반 기계학습 알고리즘을 이용한 야구 경기 Big Data 분석. *한국통신학회 학술대회논문집*, 262-265.

김주학, 조선미, 강지연(2022). 야구 경기 승패 예측을 위한 합성곱 신경망 (CNN) 최적화 연구. *한국체육측정평가학회지, 24*(4), 153-165.

김차용(2001). 프로야구경기 분석을 통한 승·패 예측 모형. *한국사회체육학회지, 16*(6), 807-819.

김태훈, 임성원, 고진광, 이재학(2020). 인공지능 모델에 따른 한국 프로야구의 승패 예측 분석에 관한 연구. *한국빅데이터학회지, 5*(2), 77-84.

김현규, 이제영(2017). 한국프로야구에서 선발투수의 투수능력지수 제안: 대체선수대비승수 (WAR) 을 중심으로. *한국데이터정보과학회지, 28*(4), 863-874.

배중현(2020년 04월 02일). KBO, 리그 기록 데이터화 완료 전준호 통산 도루 1개 감소. 일간스포츠

서영진, 문형우, 우용태(2019). 기계학습 기법을 이용한 한국프로야구 승패 예측 모델. *한국컴퓨터정보학회논문지, 24*(2), 17-24.

양도업(2016). *프로야구 연봉과 경기력 분석을 위한 세이버메트릭스 활용방안*. 미간행 박사학위논문. 고려대학교 대학원.

양도업, 조은형, 배상우, 정상원(2015). 한국 프로야구 타자의 경기력요인 분석, *한국사회체육학회지, 60*, 305-313.

양민정(2022). *수영 경기결과 예측을 위한 머신러닝 기법 비교*. 미간행 박사학위논

문. 단국대학교 대학원.

엄대엽, 김성용(2022). 머신러닝을 이용한 골든글러브 수상 요인 분석에 대한 연구. *한국콘텐츠학회논문지, 22*(5), 48-55.

예원진, 이성노(2022). 2022 FIBA 남자농구 아시안컵 경기결과를 활용한 머신러닝 분류 모형의 예측 성능 비교. *한국체육측정평가학회지, 24*(3), 53-69.

이장택, 김용태(2006). 한국프로야구에서의 승률 추정에 관한 연구. *한국자료분석학회지, 8*(2), 857-869.

이장택(2014). 한국프로야구에서 타자능력의 측정. *한국데이터정보과학회지, 25*(2), 349-356.

이장택(2014). 한국프로야구에서의 투수평가지표. *한국데이터정보과학회지, 25*(3), 485-492.

전용배(2001). 한국프로야구 규약 및 계약관계와 선수협의회에 대한 법적 고찰. *한국체육학회지, 40*(4), 515-526.

전용배, 김애랑(2011). 한,일 야구의 사회, 문화적 함의 비교. *일본근대학연구, 34*, 309-325.

조선미, 김주학, 강지연, 김상균(2023). 머신러닝 (XGBoost) 기반 미국프로야구 (MLB) 의 투구별 안타 및 홈런 예측 모델 개발. *한국체육측정평가학회지, 25*(1), 65-76.

최경호(2009). 세이버 메트릭스 소개 및 통계적 측면의 한국프로야구 기록 분석. *사회과학논총, 25*(1), 129-139.

최형준, 엄한주(2020). 스포츠경기분석의 이슈와 전망. *한국체육측정평가학회지, 22*(3), 105-113.

최형준(2022). 축구의 경기 결과 예측을 위한 머신러닝 기법 비교. *한국체육측정평가학회지, 24*(4), 81-91.

한국야구위원회. http://www.koreabaseball.com

한정섭, 정다현, 김성준(2022). 머신러닝을 활용한 빅데이터 분석을 통해 KBO 타자의 OPS 예측. *차세대융합기술학회논문지,* 6(1), 12-18.

Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152).

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers (pp. 177-186). Physica-Verlag HD.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145-1159.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Cessie, S. L., & Houwelingen, J. V. (1992). Ridge estimators in logistic regression. Journal of the Royal Statistical Society Series C: *Applied Statistics*, *41*(1), 191-201.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1-4.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., ... & Zhang, Z. (2015). Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J.,

& Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783-2792.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In Breakthroughs in statistics: Methodology and distribution (pp. 569-593). New York, NY: Springer New York.

Franks, I. M., & Goodman, D. (1986). A systematic approach to analysing sports performance. *Journal of Sports Sciences, 4*(1), 49-59.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33*(1), 1-22.

Hart, P. E., Stork, D. G., & Duda, R. O. (2000). Pattern classification. Hoboken: Wiley.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Heo, J. H., & Woo, Y. T. (2023). An Estimation Model for Defence Ability Using Big Data Analysis in Korea Baseball. *Journal of The Korea Society of Computer and Information, 28*(8), 119-126.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science, 313*(5786), 504-507.

Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

Hughes, M., & Franks, I. M. (Eds.). (2004). Notational analysis of sport: Systems for better coaching and performance in sport. Psychology Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Ji, Y., Zhanga, H., Zhangb, Z. & Liu, M. (2021). CNN-based encoder-decoder networ ks for salient object detection:A comprehensive review and recent advance s. *InformationSciences 546*, 835-857. https://doi.org/10.1016/j.ins.2020.09.003

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis, 9*(2), 137-163.

Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436-444.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

Loh, W. Y. (2011). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge *discovery, 1*(1), 14-23.

McGarry, T., ODonoghue, P., & de Eira Sampaio, A. J. (Eds.). (2013). Routledge handbook of sports performance analysis. Routledge.

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific data, 6*(1), 236-250.

SABR (2016). http://sabr.org/about

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. I*BM Journal of research and development, 44*(1.2), 206-226.

Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks, 10*(5), 988-999.

(Abstract)

Comparison of prediction models for Korean professional baseball Golden Glove winners

KWON Soongyu

Department of Physical Education

Graduate School of Dankook University

Advisor : Prof. Choi Hyongjun

Baseball, compared to other sports, has diverse and meticulously preserved records, allowing a retrospective analysis of game content and outcomes through all recorded events that occurred during the matches. Consequently, baseball is commonly referred to as the "sport of statistics." In baseball, with its abundance of various records, there exists the "Golden Glove" award given to the best player in each position. Currently, the Golden Glove is awarded through voting among media professionals such as reporters, broadcast producers, and analysts, leading to limitations in identifying clear criteria and significant variables for each position. Therefore, the purpose of this study is to design a predictive model for Golden Glove winners using logistic regression analysis and machine learning, based on records of Golden Glove candidates and winners in the Korean professional baseball league from 2003 to 2022. The designed predictive model's performance is then compared and analyzed to identify the most suitable model for predicting Golden Glove winners. Additionally, the study aims to derive variables that significantly influence awards for each position. To achieve the objectives of this research, logistic regression analysis, Support Vector Machine (SVM), Random Forest, and XGBoost models were designed, and hyperparameters for each model were presented. Z-scoring and min-max scaling were used to represent each model in two different forms, and after exploring optimal variables for each model,

performance evaluations were conducted.

Firstly, in designing the predictive model for Korean professional baseball's Golden Glove using logistic regression analysis, L1, L2, and elasticnet were used as kernels, while in the Support Vector Machine model, rbf and poly were used as kernels, and important variables were not explored as it is a non-linear model. In the Random Forest model, gini and entropy were used as criteria, and in the XGBoost model, exact, approx, and hist were used as criteria. The removal of variables was conducted one by one to improve the F1 score, and the variables selected for each position in each model were different.

Secondly, comparing the predictive performance of machine learning models revealed differences in accuracy and F1 score for each position. SVM model 2 and XGBoost model showed high accuracy and F1 scores, while logistic regression analysis and Random Forest models showed relatively lower accuracy and F1 scores. Overall, Model 2, standardized with min-max scaling, demonstrated superior predictive ability compared to Model 1, standardized with z-scoring.

In conclusion, it was possible to predict Golden Glove winners based on the records of candidates and winners, and important variables for each model were identified. Furthermore, through performance evaluation, XGBoost Model 2 appeared to be the most suitable model. Thus, it is considered desirable to predict the Golden Globe Award winner by utilizing XGBoost Model 2 with minmax. Moreover, it is believed that incorporating defensive records in future predictions will yield superior results.

keywords : Korean Professional Baseball, Sabermetrics, Golden Glove Award Predictions, Logistic Regression Analysis, Machine Learning, Prediction Techniques