

석사학위논문

의사결정나무 분석 기법을
활용한 프로야구 외국인 투수
재계약 확률 예측

(A study on KBO league foreign pitchers'
re-sign possibilities using decision tree analysis)

고려사이버대학교 융합정보대학원

융합정보학전공

황 규 인

2017년 12월

오수길 교수지도

석사학위논문

의사결정나무 분석 기법을
활용한 프로야구 외국인 투수
재계약 확률 예측

(A study on KBO league foreign pitchers'
re-sign possibilities using decision tree analysis)

이 논문을 융합정보학 전공 석사학위
논문으로 제출함.

2017년 12월

고려사이버대학교 융합정보대학원

융합정보학전공

황 규 인 (인)

황규인의 융합정보학전공 석사
학위논문 심사를 완료함

2017년 12월

위원장 백 창 현 (인)

위 원 권 오 상 (인)

위 원 오 수 길 (인)

요약

한국 프로야구는 2017년 외국인 선수 제도 도입 20주년을 맞았지만 각 구단은 여전히 외국인 투수 영입 과정에서 선수 출신 스카우트의 관찰과 직감에 의존하는 게 일반적이다. 결과는 별로 성공적이지 못하다. 2007년부터 2016년까지 10년간 한국 프로야구 무대에 첫 선을 보인 외국인 투수 130명 중 33명(25.4%)이 재계약에 성공했을 뿐이다.

반면 게임 회사를 모기업으로 두고 있는 NC는 2012년 창단 때부터 데이터팀을 별도로 만들어 외국인 선수 영입에 나섰다. 결과는 성공적이다. 외국인 투수 해커(34)는 NC가 1군 무대에 첫 진입한 2013년부터 2017년까지 5년 연속 함께 했으며, 같은 해 합류했던 찰리(32) 역시 노히트 노런을 기록하는 등 3년 동안 NC 마운드를 지켰다.

그 결과 NC는 1군 진입 2년째였던 2014년부터 포스트 시즌 진출에 성공하는 등 신생팀 한계를 극복하고 안정적으로 팀 전력을 끌어올릴 수 있었다. 소위 ‘세이버메트릭스’ 이론을 바탕으로 데이터(야구 기록)를 통해 외국인 선수와 계약하는 게 실질적으로 효과를 거둘 수 있다는 증거다.

단, NC는 구체적으로 어떤 기준으로 외국인 선수를 영입하는지는 공개하지 않고 있다. 이에 ‘타이태닉 생존자 분석’ 등에 활용하는 의사결정나무 기법을 활용해 외국인 투수의 재계약에 어떤 기록이 가장 영향을 많이 주는지 연구했다.

세이버메트릭스에서 투수를 평가할 때는 능력과 성과를 구분하는 작업부터 시작한다. 이때 능력은 해마다 기록이 크게 변하지 않는 분야를 뜻하고, 성과는 그 능력에 (행)운을 더해 나타난 결과를 의미한다.

DIPS(Defense Independent Pitching Statistics) 이론에 따르면 탈삼진, 볼넷 허용, 홈런 허용 등 세 가지만을 어떤 투수의 능력이라고 통계적으로 유의미하게 규정할 수 있다. 이런 능력은 갑자기 늘거나 줄지 않는다. 세이버메트릭스에서는 이 세 가지 기록이 실제 플레이로 이어지지 않는다는 뜻에서 BNIP(Balls Not In Play)라고 부

른다. 플레이로 이어지지 않는다는 건 투수와 타자 사이에서 승부가 끝난다는 뜻이다.

반면 투수를 평가할 때 가장 널리 쓰는 평균자책점은 이 능력에 운이 개입한 성과다. 이는 평균자책점에는 타자가 타구를 페어 지역에 때렸을 때 타율을 나타내는 BABIP(Batting Averages On Balls In Play)가 적지 않은 영향을 끼치기 때문이다. DIPS 이론에 따르면 BABIP에 가장 큰 영향을 끼치는 건 운이다. 실제로 삼진, 볼넷, 홈런 등과 달리 BABIP는 해마다 변동이 크다. 자연스레 같은 투수의 평균자책점도 해마다 큰 폭으로 변한다.

그럼에도 불구하고 평균자책점이 의사결정나무 분석 결과 프로야구 구단에서 외국인 투수와 재계약을 결정할 때 가장 큰 영향을 끼치는 요소로 나타났다. 이렇게 재계약을 맺은 투수는 한국 무대 데뷔 첫 해 평균 평균자책점 3.60으로 안정적인 면모를 선보였지만 이듬해에는 4.20으로 평균자책점이 올랐다. 평균자책점이 외국인 투수를 평가하는 데 있어 크게 믿을 만한 지표가 되지 못한다는 증거다.

재계약한 투수는 BNIP도 좋았지만 BNIP가 좋은 투수라고 꼭 재계약하는 건 아니었다. 한국 무대에서 미국 마이너리그 AAA로 돌아간 선수들 성적을 보면 한국 무대에서 BNIP 특히 삼진 기록이 좋았던 선수가 좋은 성적을 나타냈다. BNIP가 능력이라는 또 한 가지 증거가 될 수 있다.

좋은 외국인 투수는 한국 무대에서 오래 활약할 가능성이 높다. 의사결정나무 분석 모델을 통해 한국에서 뛴 총 기간을 예측하면 탈삼진 대 볼넷 비율이 높은 선수가 한국 무대 활동 기간이 길었다는 사실을 확인할 수 있다.

‘삼진 ÷ 볼넷’으로 계산하는 이 비율을 높이려면 삼진을 많이 잡거나 볼넷을 적게 내줘야 한다. 마이너리그에서 한국 프로야구로 건너올 때 기록 변화를 측정하면 삼진 변화량이 이 비율을 조금 더 잘 설명한다는 사실을 알 수 있다. 이는 삼진 변화가 해당 외국인 투수가 한국 프로야구 무대에서 성공할 수 있는지 따져볼 수 있는 가늠쇠로 기능할 수 있다는 사실을 나타낸다.

의사결정나무 분석을 통해 마이너리그에서보다 한국에서 삼진이 많은 투수는 어떤 구종을 잘 던지는 알아보면 커브가 답이라는 사실을 확인할 수 있다. 커브는 외국인 투수 재계약에도 유의미한 영향을 끼치는 구종이었다.

외국인 선수의 키 역시 한국 프로야구에서 해당 투수가 얼마나 오래 살아남을 수 있을지를 알려주는 변수였지만 BABIP+를 제외하면 키에 따른 변화를 알기는 어려웠다. 해외 연구 결과에 따르면 BABIP는 투수 키와 무관하다. 단, 구종 정보와 결합한 연구는 아직 찾아보기가 힘든 만큼 앞으로 이에 대한 추가 연구가 필요하다고 할 수 있다.

주제어: 프로야구, 세이버메트릭스, 기계학습, 의사결정나무, 랜덤포레스트

차례

I. 서론	1
1. 연구 목적 및 필요성	1
2. 연구 문제	3
II. 이론적·제도적 배경	3
1. 한·미 프로야구 제도 및 구조	3
2. 세이버메트릭스와 기록 예측	9
3. 투구 추적 시스템	19
4. 의사결정나무와 확률	25
III. 연구 방법	29
1. 연구 대상	29
2. 연구 절차	30
IV. 연구 결과	37
1. 외국인 투수 재계약 예측 의사결정나무 모델	37
2. BNIP 대(對) BIP	39
3. 삼진, BABIP 그리고 재계약	41
4. 투수 유형별 마이너리그 복귀 성적 비교	42
5. 외국인 투수 총 시즌 예측 의사결정나무 모델	45
6. 삼진 대 볼넷 변화 비교	46
7. 외국인 투수 구종 가치 의사결정나무 모델	48
V. 결론	50
1. 한국 커브와 미국 커브는 다르다	50
2. 키 큰 투수는 BABIP를 억제할까	52
2. 향후 연구 제안	53
VI. 참고문헌	56
VII. 부록	58

표 차례

〈표 1〉 2017년 9월 현재 프로야구 10개 구단 외국인 선수 포지션	4
〈표 2〉 2016년 기준 마이너리그 시스템	7
〈표 3〉 2012~2014년 한국 프로야구 특정 볼카운트 이후 타격 기록	23
〈표 4〉 2017년 한국 프로야구 타격 결과별 득점 가치	24
〈표 5〉 외국인 투수 데이터 파일 구조	31
〈표 6〉 마이너리그 기록 데이터 파일 구조	33
〈표 7〉 리그 기록 데이터 파일 구조	34
〈표 8〉 재계약 여부에 따른 외국인 투수 BNIP 기록 비교	39
〈표 9〉 투수 유형별 특징	43
〈표 10〉 투수 유형별 AAA 복귀 후 평균자책점+	44
〈표 11〉 투수 유형별 AAA 복귀 후 9이닝당 삼진+	44
〈표 12〉 K/BB+ 변화량 회귀 분석 결과	47
〈표 13〉 K+ 향상 여부에 따른 커브 구종 가치/100	50
〈표 14〉 2014~2016년 외국인 투수 커브 구종 가치/100	50
〈표 15〉 키 189cm 기준 BABIP 차이	52
〈표 16〉 연구 대상 외국인 투수 명단	58

그림 차례

[그림 1] 리그별 연도별 평균 득점 변화	8
[그림 2] 타석 결과를 트리 다이어그램	13
[그림 3] 연속 두 시즌 삼진 비율 상관관계	14
[그림 4] AAA에서 메이저리그 승격 때 BNIP 비율 변화	17
[그림 5] PFX 활용 메이저리그 중계 화면	20
[그림 6] 트랙맨 활용 메이저리그 TV(MLB.tv) 중계 화면	21
[그림 7] 의사결정나무 기법 활용 타이태닉 생존자 예측 모델	26
[그림 8] 랜덤 포레스트 구조	29
[그림 9] 외국인 투수 파일 스크린샷	32
[그림 10] bbref 리그 페이지 스크린샷	33
[그림 11] 스탯티즈 구종 가치 페이지 스크린샷	35
[그림 12] 외국인 투수 재계약 여부 의사결정나무 모델	37
[그림 13] 외국인 투수 평균자책점과 BABIP 상관관계	38
[그림 14] 외국인 투수 재계약 여부 의사결정나무 모델 가지치기	40
[그림 15] 외국인 투수 K-BABIP 분류	42
[그림 16] 외국인 투수 총 시즌 예측 의사결정나무 모델	46
[그림 17] 외국인 투수 재계약 확률 예측 랜덤 포레스트 모델 변수 중요도	48
[그림 18] 리그 변화에 따른 K+ 향상 예측 의사결정나무 모델	49
[그림 19] 한국 프로야구와 AAA 커브 릴리스 각도 차이	51

I. 서론

1. 연구 목적 및 필요성

외국인 투수가 한국 프로야구 데뷔 첫 해 남긴 기록을 토대로 이 선수가 한국 무대에서 얼마나 오래 활약할 수 있는지 예측할 수 있는 모델을 설계하는 게 이번 연구 목적이다. 이를 통해 프로야구 각 구단에서 외국인 투수에게 투자하는 과정에서 낭비를 줄일 수 있도록 도움을 주는 데 이번 연구 의의가 있다고 할 수 있다.

한국 프로야구는 1982년 출범 이후 1997년까지 16년간 한국 출신 및 재일동포 선수들에게만 출전을 허용했다. 그 이후 1998년 외국인 선수 제도를 도입하면서 2016년까지 외국인 선수 총 323명이 한국 프로야구 무대를 거쳐 갔다.

이 중 64.4%(208명)가 투수일 정도로 각 프로야구 구단은 외국인 선수로 타자(야수)보다는 투수를 선호했다. 기본적으로 ‘야구는 투수 놀음’이라는 말이 웅변하는 것처럼 투수가 전력에서 차지하는 비중이 높다는 게 첫 번째 이유다. 게다가 돌아가며 마운드에 오르는 투수와 달리 타자는 거의 매일 경기에 출장하기 때문에 국내 선수 자리를 빼앗는다는 문제가 있다.

이 때문에 각 프로구팀에서는 외국인 선수로 투수를 선호한다. 갈수록 이 같은 투수 현상은 심해져 2017년 개막전에는 10개 팀 모두 외국인 투수를 선발투수로 내세울 만큼 프로야구 무대에서 외국인 선수가 차지하는 비중이 올랐다.

2014년 한국야구위원회(KBO)에서 ‘야구 규약’에 나와 있던 외국인 선수 몸값 상한선을 폐지하면서 이들이 실제로 받고 있는 돈이 얼마인지 공개되기 시작했다. 이전에는 야구 규약상 최고액인 30만 달러(약 3억 3600만 원)로 발표하는 게 일반적이었다.

실제 2017년 개막일 기준으로 10개 구단이 외국인 투수 20명에게 지급한 몸값(계약금+연봉)은 총 2205만5000 달러(약 247억 160만 원)까지 올랐다. 선수당 평균 금액은 100만 달러가 넘어 2014년 이전 규정보다 3.6배 이상 많았다.

각 구단은 야구 규약에 따라 외국인 선수에게 연봉뿐 아니라 선수와 그 가족

에게 구단 돈으로 숙소와 교통, 통역도 제공해야 한다. 물론 외국인 선수 영입에 쓰는 스카우트 비용도 무시 못 할 수준이다.

하지만 외국인 선수 투자가 아주 큰 결실을 맺었다고 보기는 힘들다. 최근 10년(2007~2016년) 한국 프로야구 무대에 첫 선을 보인 외국인 투수 130명¹⁾ 가운데 재계약에 성공한 건 33명(25.4%)밖에 되지 않는다.

반면 프로야구 제9구단 NC는 1군 무대에 뛰어든 처음 뛰어든 2013년 영입한 해커(34)와 네 시즌 연속 재계약했고, 같은 해 처음 영입한 찰리(32)도 세 시즌 동안 활약했다. 찰리는 6월 24일 LG를 상대로 노히트 노런(투수가 안타를 하나도 맞지 않고 점수도 한 점도 내주지 않는 일)을 기록하기도 했다.

대부분 구단은 선수 출신 스카우트들이 자기 경험을 바탕으로 관찰과 직감에 의존해 외국인을 뽑는다. 이와 달리 게임 회사를 모기업으로 두고 있는 NC는 비(非)선수 출신인 야구 통계 전문가들이 모여 데이터를 보고 외국인을 데려온다(조선일보, 2017).

결국 데이터(야구 통계) 기반으로 외국인 선수를 영입하는 게 실질적으로도 가능한 일이지만 지금까지는 이에 관한 연구가 거의 없는 실정이다. NC 역시 ‘영업 기밀’을 이유로 외국인 선수 선발 기준을 외부에 공개하지 않는다. 따라서 기록을 토대로 국내 무대에서 생존 가능한 외국인 선수를 예측할 수 있는 방안이 무엇인지 관한 연구가 필요하다고 할 수 있다.

이런 연구가 결실을 맺게 되면 만성적자에 시달리고 있는 각 프로야구 구단의 투자 효율화에 도움을 줘 프로야구 산업 발전에도 기여할 수 있을 것이다. 한국 프로야구에 관한 기본적 시각은 대기업 주요 계열사가 비용을 부담해 의미 없는 사업을 사회사업 관점에서 진행한다고 인식된다는 점이다. 이는 1982년 프로야구 창설 자체가 경제논리보다는 정치권 주도로 진행되었기 때문이며, 여전히 각 기업들의 수익구조가 좋지 않기 때문이기도 하다(정호윤·이상우, 2015) 한국 프로야구 소속

1) 한국 프로야구 무대 첫 등장 기준. 국내 팀 A에서 활약한 뒤 나중에 팀 B와 재계약하는 경우는 제외

10개 팀은 모기업이 없는 넥센을 제외하면 대부분 매출액 40% 이상을 모기업에 의존하고 있는 실정이다(일간스포츠, 2017).

사정이 이런데도 한국 프로야구 구단이 검증되지 않은 선수에게 일본 프로야구보다 더 많은 금액을 지불하고 있다는 지적(일간스포츠, 2016a)이 나오고 있다. ‘가격 대비 성능 비율’이 보다 뛰어난 외국인 투수를 선발할 수 있는 방안에 대한 연구가 필요한 이유다.

2. 연구 문제

이번 연구에서는 이미 한국 프로야구를 경험한 외국인 선수가 국내 프로야구 무대에서 남긴 기록을 토대로 재계약 가능성을 예측할 수 있는지 살펴볼 것이다. 이를 토대로 해외 리그 기록을 토대로 이런 기록을 예측할 수 있는지 알아보는 것 역시 이번 연구 목표다. 이에 따른 이번 연구 문제는 다음과 같다.

- 한국 프로야구 첫 시즌 성적으로 재계약 여부를 예상할 수 있는가
- 그렇다면 어떤 기록이 외국인 투수 재계약에 가장 큰 영향을 끼치는가
- 정말 이 기록이 외국인 투수 재계약에 가장 큰 영향을 끼쳐야 하는가
- 그 기록에 도움을 주는 투수 요소는 무엇인가

Ⅲ. 이론적 · 제도적 배경

1. 한미 프로야구 제도 및 구조

가. 한국 프로야구 외국인 선수 제도

KBO는 2014년부터 구단별로 외국인 선수를 3명까지 보유할 수 있도록 규정을

손질했다. 단, 선수 포지션을 투수와 야수(타자)로 나뉘었을 때 두 포지션에서 최소한 명을 뽑아야 한다. 그러니까 ‘투수 2명 + 타자 1명’ 또는 ‘투수 1명 + 타자 2명’ 조합을 선택해야 하는 것이다.

또 한 경기에 외국인 최대 2명까지만 출전할 수 있기 때문에 각 구단은 ‘선발 투수’ 두 명과 타자 한 명을 뽑는 게 일반적이다(<표1 참조>). 만약 외국인 선수가 모두 타자라면 외국인 투수가 선발 투수일 때 한 명을 출전시킬 수 없어 전력 누수가 뒤따른다. 외국인 투수가 구원 투수일 때도 문제다. △선발 투수가 외국인이고 △이미 외국인 야수가 출전한 상황이라면 구원 투수가 경기에 나설 수 없기 때문이다.

<표 1> 2017년 9월 현재 프로야구 10개 구단 외국인 선수 포지션

구단	이름	포지션	구단	이름	포지션
넥센	밴헤켄	투수	KIA	버나디나	외야수
	브리검	투수		팻딘	투수
	초이스	외야수		헥터	투수
두산	니퍼트	투수	kt	로치	투수
	보우덴	투수		로하스	외야수
	에반스	외야수		피어밴드	투수
롯데	레일리	투수	LG	소사	투수
	린드블럼	투수		허프	투수
	번즈	내야수		로니 ²⁾	내야수
삼성	러프	내야수	NC	맨심	투수
	레나도	투수		스크릭스	내야수
	페트릭	투수		해커	투수
한화	로사리오	포수	SK	다이아몬드	투수
	비야누에바	투수		로맥	외야수
	오간도	투수		켈리	투수

2) 구단에서 퓨처스리그(2군)행을 통보하자 2017년 8월 27일 무단귀국으로 임의탈퇴

2003년부터 2013년까지는 팀당 외국인 선수 보유 한도가 2명이었다. 이때는 외국인 타자 없이 외국인 투수 두 명을 뽑는 게 일반적이었다.

예외적으로 2013년 1군 무대에 뛰어든 NC는 신생팀 보호 차원에서 2013년에는 3명, 2014년에는 4명까지 외국인 선수를 보유할 수 있었다. 2015년 1군 무대에 진입한 kt는 그해와 2016년에 외국인 선수를 4명 보유했다.

2017년 한국 프로야구 구단과 계약한 외국인 투수 중에서는 두산에서 일곱 번째 시즌을 맞이한 니퍼트(36)가 210만 달러(약 23억 5200만 원)로 최고 몸값을 받는다. 니퍼트가 메이저리그에서 뛰었던 마지막 시즌(2010년) 미국에서 연봉 66만5000달러를 받았다는 점을 감안하면 한국 구단과 계약을 연장하는 게 외국인 투수에게도 도움이 되는 일이라고 짐작할 수 있다.

나. 미국 프로야구 마이너리그 시스템

국립국어원 ‘표준국어대사전’은 ‘프로야구’를 “직업 선수들이 생활의 수단으로 하는 야구 즉 직업 야구”라고 풀이한다. 한국에서는 KBO가 주관하는 리그를 프로야구라고 부르지만 미국에는 관점에 따라 프로야구로 볼 수 있는 여러 야구 리그가 존재한다.

이 중 가장 대표적인 존재가 내셔널리그(NL)와 아메리칸리그(AL)라는 양대 ‘메이저리그(major leagues)’를 축으로 하는 ‘조직 야구(organized baseball)’ 시스템이라고 할 수 있다.

NL과 AL을 다른 리그와 분류해 메이저리그라고 부르는 것부터가 미국에 다양한 프로야구 리그가 존재하고 있다는 사실을 증명한다. 메이저리그를 축으로 하는 야구 시스템을 조직 야구라고 부르는 건 이 조직에 속하지 않은 독립 리그가 존재하기 때문이다.

메이저리그에는 양대 리그에 각 15개 팀씩 총 30개 팀이 속해 있으며 각 팀은 ‘팜(farm) 시스템’이라고 부르는 선수 육성 시스템을 갖추고 있다. 이 팜 시스템

을 마이너리그(minor leagues)라고 부른다. 마이너리그 팀은 메이저리그 팀과는 독립적인 별개 법인이며 메이저리그 팀과 선수 위탁 교육 계약을 맺는다.³⁾

연령 등으로 인해 선수 수준에 차이가 나는 것처럼 자연스레 마이너리그 팀 역시 서로 다른 수준을 유지하고 있다. 마이너리그는 선수 수준이 낮은 리그부터 높은 리그 순서로 R(루키) - A(싱글 A) - AA(더블 A) - AAA(트리플 A)로 크게 구분할 수 있다. 이 명칭은 리그 수준(level)을 표시한 것으로 리그 자체 명칭은 아니다.⁴⁾

예컨대 AAA는 크게 인터내셔널리그(IL·International League)와 퍼시픽코스트 리그(PCL·Pacific Coast League)로 나뉘어 있다. 또 원래 멕시코 프로야구 리그인 ‘멕시코리그’는 조직 야구로 묶인 리그는 아니지만 주로 AAA 팀과 선수 교류를 하는데다 메이저리그 선수를 많이 배출해 보통 AAA 수준으로 간주한다.

조직 야구에서 몸담게 된 모든 선수가 이 단계를 차례대로 거쳐 메이저리그로 올라가는 건 아니다. 유망주 가운데는 AAA를 거치지 않고 AA에서 곧바로 메이저리그로 올라가는 선수도 적지 않다. 이는 기본적으로 AAA 소속 팀은 선수 육성 기능뿐 아니라 이름 그대로 메이저리그에 가기 힘든 선수들이 뛰는 ‘하위(마이너) 리그’라는 성격도 있기 때문이다.

한국 프로야구에 오는 외국인 선수들은 한국 팀과 계약하기 전 AAA에서 뛰고 있던 게 일반적이다. 거꾸로 일정 시간이 지나도록 AAA에 올라갈 실력이 되지 못하는 선수는 조직 야구를 떠나야 하는 상황에 놓이게 돼 한국 무대를 밟기도 쉽지 않다. 이런 이유로 이번 연구에서는 ‘해외 리그’를 AAA로 특정하도록 한다.

3) 이와 비교하면 한국 프로야구 2군 팀은 같은 법인 내에 다른 ‘부서’라고 할 수 있다.

4) 원래 이 수준은 메이저리그 팀에서 각 리그 소속 선수 계약을 ‘사올 때’ 계약금 표준 규정을 만들려는 목적으로 나눈 것이다. 예전에는 메이저리그 팀에서 마이너리그 팀에 선수 육성을 위탁하는 방식이 아니라 마이너리그에서 선수를 ‘발견’하면 계약을 인수하는 형식이 메이저리그-마이너리그 관계였기 때문이다. 현재 방식처럼 기량 발전 속도에 따라 선수를 파견하게 된 건 브랜치 리키 세인트루이스 단장이 1928년부터 선수가 아니라 아예 구단을 사들이기 시작한 이후다. ‘일반적으로 자회사가 아닌’ 이유는 이처럼 메이저리그 구단에서 마이너리그 팀을 보유하고 있는 사례도 있기 때문이다.

〈표 2〉 2016년 기준 마이너리그 시스템

리그 수준	리그	소속팀 숫자	경기당 득점
AAA	인터내셔널리그	14	3.96
	퍼시픽코스트리그	16	4.78
	멕시코리그	16	4.71
AA	이스턴리그	12	4.34
	서던리그	10	4.08
	텍사스리그	8	4.12
A 어드밴스트	캘리포니아리그	8	5.01
	캐롤라이나리그	10	4.52
	플로리다스테이트리그	12	4.04
A	미드웨스트리그	16	4.05
	사우스애틀랜틱리그	14	4.20
단기 시즌 A	뉴욕펜리그	14	4.03
	노스웨스트리그	8	4.82
R	파이니어리그	8	5.91
	애플랜치안리그	10	4.85
	애리조나리그	14	5.00
	걸프코스트리그	17	4.12

다. 기록+(플러스)

레벨이 같은 리그라고 해도 경기 진행 양상은 서로 다를 수 있다. 예컨대 어떤 리그가 다른 리그보다 ‘타자친화적’이라 점수가 더 많이 난다면 투수를 평가하는 기준도 달라져야 한다. 한 팀이 평균 4.0 득점하는 리그에서 뛰는 투수라면 평균자책점⁵⁾ 4.50은 평균보다 못한 투수지만, 5.0점을 내는 리그라면 평균보다 뛰어난 투수이기 때문이다.

또 같은 리그라고 해도 [그림 1]을 통해 확인할 수 있는 것처럼 해마다 득점 환경이 변하는 게 일반적이다.

5) 예전에는 ‘방어율’이라고 부르던 기록. 현재는 공식 ‘야구 규칙’에도 ‘평균자책점(방어율)’이라고 표기한다.



[그림 1] 리그별 연도별 평균 득점 변화

이럴 때는 ‘평균자책점+’ 처럼 기록+를 활용해 수준 차이를 보정하는 일이 많다. 기록+는 기본적으로 아래 같은 공식을 통해 계산한다.

$$\text{기록+} = \frac{\text{개인 기록}}{\text{리그 기록}} \times 100$$

말하자면 한 선수 기록이 리그 평균보다 얼마나 좋은가 혹은 얼마나 나쁜가를 백분율로 나타내는 기록이 바로 기록+인 셈이다.

이 기록이 높다고 무조건 좋은 건 아니다. 예를 들어 타율처럼 원래 높을수록 좋은 기록은 기록+도 높으면 좋다. 반면 평균자책점처럼 낮아야 좋은 기록은 기록+ 역시 낮을수록 좋다.

이렇게 리그와 시대 차이를 보정하는 기록+는 한국 언론에서 흔히 ‘야구 통계학’ 이라고 번역하는 ‘세이버메트릭스’ 이론의 가장 기본 단계라고 할 수 있다.

2. 세이버메트릭스와 기록 예측

가. 세이버메트릭스의 이해

‘세이버메트릭스(Sabermetrics)’는 아직 표준국어대사전에 실리지 않은 낱말이다. 이 영어 낱말은 SABRmetrics를 소리 나는 대로 표기한 것이다. SABR에서 쓰는 지표(metric)가 바로 세이버메트릭스다. SABR은 ‘Society for American Baseball Research(미국야구연구학회)’를 줄인 말이다. 결국 야구를 (과학적으로) 연구해 세상에 내놓는 학문이 바로 세이버메트릭스다.

따라서 꼭 야구 통계를 다루는 분야만이 세이버메트릭스인 것은 아니다. 하지만 일반적으로는 야구 관계자들이 경험을 통해 얻은 ‘직관’이 아니라 통계적인 접근을 통해 야구를 바라보는 관점을 흔히 세이버메트릭스라고 부른다. 그런 이유로 많은 이들이 세이버메트릭스를 야구 통계학과 동의어로 받아들인다.

세이버메트릭스라는 낱말은 경제 저널리스트 마이클 루이스가 메이저리그 오 클랜드 어슬레틱스의 성공 사례를 분석해 2003년 세상에 내놓은 책 ‘머니볼: 불공정한 게임을 승리로 이끄는 과학(Moneyball: The Art of Winning an Unfair Game)’을 통해 대중에게 알려지기 시작했고, 이 기법을 도입한 보스턴 레드삭스가 2004년 월드시리즈에서 우승하면서 86년 만에 메이저리그 정상에 서자 세이버메트릭스의 가치가 더욱 올랐다.

보스턴이 ‘밤비노의 저주’⁶⁾를 깨뜨리는 데 일등공신으로 활약한 인물이 바로 ‘세이버메트릭스의 아버지’⁷⁾로 불리는 빌 제임스(68)였다. 보스턴은 2003년 그를 구단주 자문 위원으로 영입하면서 전통적인 선수 스카우트 방식과 작별을 고했다. 제임스는 이 공로를 인정받아 2006년 미국 시사주간지 ‘타임’가 성정한 세계에서 가장 영향력 있는 인물 100인으로 뽑히기도 했다. 보스턴은 그후 2007년과 2013년에도 우승을 차지하며 변화를 선택한 게 옳았다는 사실을 증명했다.

6) Curse of Bambino. 보스턴이 1920년 전설적인 홈런왕 베이스 루스를 뉴욕 양키스로 트레이드 시킨 후 우승하지 못했던 징크스를 일컫는 표현.

7) 1971년 창설 때부터 SABR 회원으로 참가했던 그는 세이버메트릭스 기본이 되는 여러 이론을 만들어낸 건 물론이고 ‘SABRmetrics’라는 표현을 만든 주인공이기도 하다.

보스턴만 이런 변화를 선택한 건 아니다. ‘머니볼’ 출간 이후 메이저리그 구단들은 세이버메트릭스 전담부서나 통계분석팀을 신설했고, 현재⁸⁾는 75%가 넘는 구단들이 최소 한 명 이상의 통계분석원을 두고 있다. 물론 세이버메트리션⁹⁾을 여러 명 고용한 구단도 많다(Baumer, B., Zimbalist, A., 2013)

나. 세이버메트릭스 선수 평가 방식

‘머니볼’에서 오클랜드가 성공한 비결로 꼽은 건 ‘출루율’이었다. 출루율은 아래 공식을 통해 계산한다.

$$\text{출루율} = \frac{(\text{안타} + \text{볼넷} + \text{몸에 맞는 공})}{(\text{타수} + \text{볼넷} + \text{몸에 맞는 공} + \text{희생플라이})}$$

‘머니볼’ 이전까지는 타자를 평가할 때 흔히 ‘5툴’이라고 부르는 △(타격) 정확도 △장타력 △주루 스피드 △(수비) 순발력 △송구 능력 등 다섯 가지를 활용하는 게 일반적이었다. 신체 능력으로 선수를 평가했던 것이다. 야구에 필요한 신체 능력이 뛰어난 선수가 야구를 더 잘할 것이라는 가정에 기반한 평가 시스템이었다.

그런 점에서 출루율은 5툴과 썩 맞아 떨어지는 기록은 아니다. 그래서 많은 이들에게 관심을 끌지 못했고 그 탓에 ‘숨어 있었다.’ 출루율을 통한 선수 평가가 성공하자 ‘숨은 기술(Hidden Art)’이라는 평가가 등장한 이유다.

출루율은 신체 능력 그 자체를 보여주지는 않지만 그 타자가 어떤 이유로든 얼마나 아웃을 당하지 않는지를 알려준다. 타석에 들어선 타자는 출루에 성공하거나 아니면 아웃을 당한 채 더그아웃으로 돌아가거나 두 가지 중 하나인 결론을 얻게 된다. 그런 이유로 오클랜드는 야구에서 공격을 ‘아웃’이라는 자원을 사용해

8) 2015년

9) Sabermetrician, 세이버메트릭스 분석가

‘득점’이라는 결과물을 만들어내는 과정이라고 규정했다. 출루율이 높은 선수는 자원을 적게 쓰는 선수가 되는 셈이다.

출루율뿐만이 아니다. 이로부터 발전한 세이버메트릭스 이론은 이렇게 모든 기록을 득점으로 환산해 평가하는 게 기본이다. 출루율은 아웃이라는 자원을 효율적으로 사용할 줄 아는 능력을 보여주기에 의미가 있다. 따라서 세이버메트릭스에서 타자를 평가할 때 중요한 건 실제로 득점을 올리는 데 필요한 선수 특성이다. 그 결과는 기록으로 확인할 수 있다.

이는 사실 ‘야구 규칙’이 지향하고 있는 관점과도 부합하다. 야구 규칙 1.02는 “각 팀의 목적은 상대팀보다 많이 득점하여 승리하는 데에 있다”고 규정하고 있다. 이 문장이 스스로 설명하고 있는 것처럼 득점이 쌓이면 한 경기에서는 승리가 되고, 한 시즌으로 보면 더 많은 승리 그러니까 더 높은 승률이 된다.

거꾸로 투수 관점에서는 ‘상대팀보다 적게 실점해 승리하는 데’ 그 투수 목적이 있다고 할 수 있다. 각 프로야구 팀에서 기량이 뛰어난 외국인 투수를 영입하려는 제일 큰 이유는 바로 이 실점을 줄이려는 목적이다.

다. DIPS(Defense Independent Pitching Statistics)

그렇다고 세이버메트릭스가 아무 기록이나 활용하는 건 아니다. 세이버메트릭스 발전 과정에서 등장한 가장 중요한 이정표 중 하나다 ‘수비 독립 투수 기록(Defense Independent Pitching Statistics)’ 그러니까 DIPS의 발견이다. 1999년 당시 대학원생이었던 보로스 매크라켄은 야구에서 투수와 수비를 분리시키는 DIPS를 고안해 내면서 투수 평가 시스템을 바꿔놓았다.

원래 야구에서 투수 능력을 기록으로 평가할 때는 승리와 패배, 평균자책점 같은 기록을 활용하는 게 일반적이었다. 승패가 운(運)에 좌우된다는 건 많은 이들이 생각했던 내용이었다. 잘 던지고도 타선에서 점수를 뽑아주지 못해 승리 투수가 되지 못하거나 심지어 패전 투수가 되는 일이 흔했기 때문이다.

매크라켄은 실점도 운에 좌우된다고 주장했다. ‘투수는 실점을 막아내는 사람’이라는 편견을 무너뜨린 것이다. 좀 더 정확하게 말하면 매크라켄은 안타를 맞아 점수를 내주는 게 거의 대부분 운에 달렸다고 주장했다(McCracken, 2001). ‘좋은 투수는 안타를 적게 맞는다’는 상식이 틀렸다는 얘기였다.

라. BABIP(Batting Averages On Balls In Play)

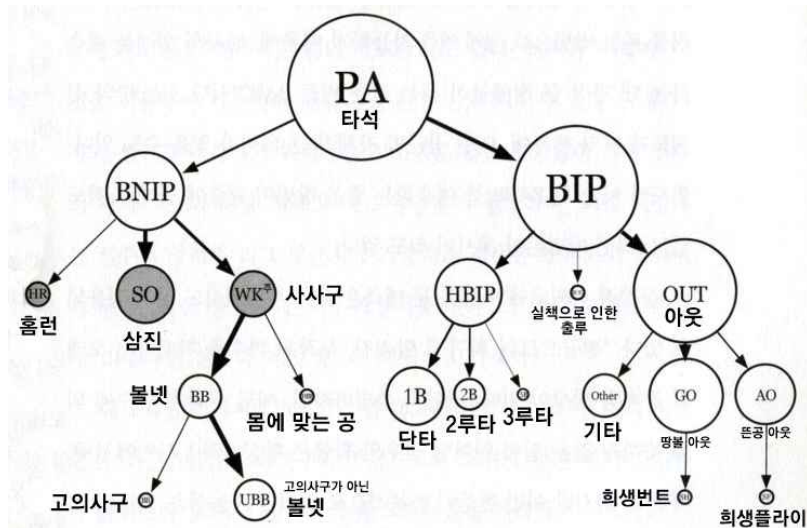
매크라켄이 근거로 삼은 건 타자가 때린 페어 타구가 안타가 되는 확률을 나타내는 BABIP(Batting Averages On Balls In Play)였다. 매크라켄은 이 BABIP가 투수에 따라 특정한 경향을 나타내지 않는다는 사실을 발견했다. 아주 좋은 투수가 던진 공이든 C급 투수가 던진 공이든 일단 타자가 공을 쳐서 담장 안 페어지역에 떨어진다면 안타가 될 확률에 큰 차이가 없다는 것이다. 매크라켄은 이를 토대로 “투수에게는 BABIP를 제어할 능력이 없다”고 주장했다.

BABIP가 일반적으로 안타가 될 확률을 나타내는 타율(안타÷타수)과 구분되는 건 홈런과 삼진이 (경우에 따라 희생플라이도) 계산에서 빠진다는 점이다.

$$BABIP = \frac{\text{안타} - \text{홈런}}{\text{타수} - \text{홈런} - \text{삼진} (- \text{희생플라이})}$$

홈런과 삼진은 야수가 개입할 여지없이 투수와 타자 사이에서 승부가 끝난다. 볼넷도 마찬가지다. 매크라켄은 이 세 경우를 ‘BNIP(Balls Not In Play) 상황’이라고 이름 붙였다.

그다음 투수는 BNIP 상황에서 벌어지는 일을 투수가 제어할 능력이 없다고 결론 내렸다. 대신 전체 투구에서 BIP(Balls In Play)가 되는 비율에는 큰 차이가 있고, 이것이 곧 투수의 능력을 반영한다고 주장했다. 요컨대 투수를 평가할 때는 삼진, 볼넷, 피홈런에 따라 평가해야 한다는 것이다.



[그림 2] 타석 결과 트리 다이어그램(Baumer, B., Zimbalist, A., 2013)¹⁰⁾

매크라켄이 초기에 세운 가설은 현재 많은 논쟁으로 힘을 잃은 상태지만¹¹⁾ 여전히 BABIP에 대해 투수가 할 수 있는 게 적다는 믿음은 세이버메트릭션을 비롯한 야구계 전반이 인정하는 사실이다(Baumer, B., Zimbalist, A., 2013).

마. 기록 예측 가능성

매크라켄이 BNIP 상황에서만 투수 능력을 검증할 수 있다고 내린 근거는 ‘예측(가능)성’이었다. 예측 가능한 것만 능력이라고 파악했던 것이다.

우리는 대체로 늘 비슷하게 유지되는 어떤 것을 그 사람의 특성이라 여긴다. 그런 까닭에 어떤 통계를 선수에게 적용했을 때 꾸준한 수치를 보여준다면, 그 통계에 의해 측정된 값을 그 선수의 특성이라 할 수 있는 근거가 된다. 반면 짧은 기

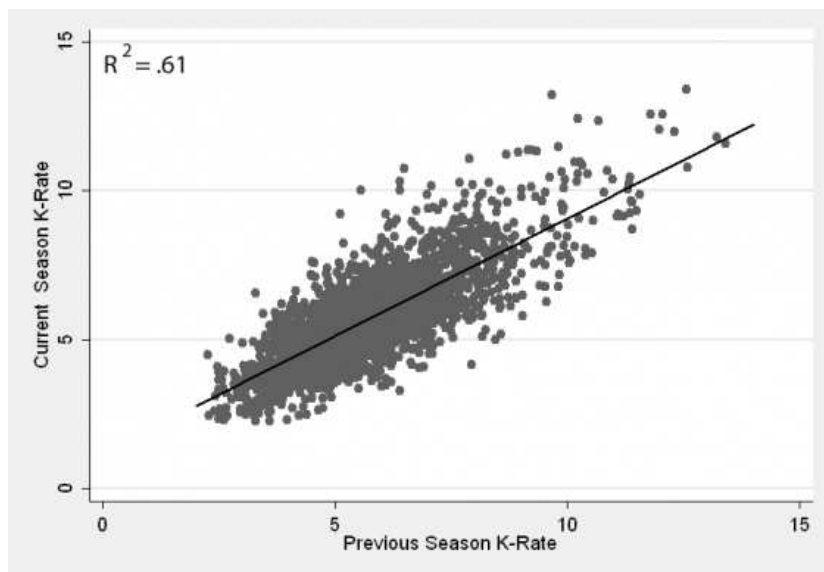
10) 원본에는 한글 용어 설명이 없다. 이번 연구용으로 덧붙인 것.

11) 이 주장이 100% 틀렸다는 의미는 아니다. 이후 연구를 통해 팀 수비 실력, (기후, 구장 모양, 외야 담장 높이 같은) 안방 구장 특성도 BABIP에 영향을 줄 수 있다는 사실이 밝혀졌다. 이 모든 요소 역시 투수 실력과 큰 관계가 없다(Baumer, B., Zimbalist, A., 2013).

간 동안에도 계속해서 불규칙적인 수치가 나온다면, 이것은 선수의 특징이라 볼 수 없으며 ‘예측성’이 낮은 특성으로 간주된다. 신뢰도를 양적으로 분석하는 방법 중 하나는 바로 통계적인 ‘자기상관(autocorrelation)’을 조사하는 것이다. 쉽게 말해, 이전의 수치와 현재의 수치가 어느 정도 연관성을 가지는지 조사한다는 이야기다(Baumer, B., Zimbalist, A., 2013).

야구에서는 보통 연도별 상관계수로 자기상관을 측정한다. 예를 들어 홈런 타자는 홈런 타자다. 이는 지난해에 홈런을 많이 친 타자가 보통 올해도 홈런을 많이 친다는 뜻이다. 투수 쪽에서는 삼진이 이런 특성을 나타내는 대표적인 기록이다.

실제 연구 결과도 그렇다. 1980~2004년을 기준으로 메이저리그에서 2년 연속 100이닝 이상을 소화한 투수의 n 년(x 축)과 $n+1$ 년(y 축) 탈삼진 비율 사이에 선형 회귀분석을 실시하면 [그림 3]과 같은 결과가 나온다.



[그림 3] 연속 두 시즌 삼진 비율 상관관계(Bradbury, 2015)

R^2 이 0.61이라는 건 전년도 삼진 비율을 토대로 올해 삼진 비율 변화를 61%

설명할 수 있다는 뜻이다. 그만큼 예측 가능성도 올라간다.

타자를 삼진으로 돌려세우는 것은 전적으로 투수의 능력에 달린 일이고, 시간이 지남에 따라 삼진 능력이 떨어진다면 한들 바로 다음 시즌에 갑자기 삼진율이 급격히 떨어지는 경우는 흔치 않다(Baumer, B., Zimbalist, A., 2013).

다른 기록도 같은 방식으로 살펴볼 수 있다. 1995~2011년 메이저리그에서 활약하며 해마다 최소 타자 250명을 상대한 투수 3077명의 n년과 n+1년 사이 ‘피어슨 상관 계수’를 구하면 0.759가 나온다. 볼넷은 0.629고, 피홈런은 0.318이다. 반면 BABIP는 상관계수가 0.171밖에 되지 않는다.

매크라켄을 이런 내용을 근거로 “결국 투수의 BABIP에 대한 가장 좋은 예측치는 리그 평균 BABIP”라고 주장하기도 했다. 따라서 투수 성적을 예측할 때는 BIP보다 BNIP 상황을 염두에 두는 게 보다 합리적이다.

이 발견이 의미가 있는 건 ‘성과’와 ‘능력’을 구분했기 때문이다. 운이 좋아서 좋은 성과를 낸 경우도 그 자체로 칭찬받을 만한 일이다. 하지만 이를 근거로 이 선수가 다음 시즌에도 좋은 성적을 거둘 것이라고 기대하기는 무리다. 반대로 운이 나빠 좋지 못한 성과를 거둔 경우에도 능력을 유지할 수 있다면 성적 반등을 기대할 수 있다.

바. FIP(Fielding Independent Pitching)

톰 탱고라는 세이버메트릭션은 이 이론을 발전시켜 BNIP를 토대로 투수 성적을 평균자책점 범위로 나타낼 수 있는 FIP(Fielding Independent Pitching · ‘수비 독립 투구’라는 뜻으로 DIPS와 의미가 같다)라는 메트릭을 개발했다. FIP는 다음과 같은 공식을 통해 계산한다.

$$FIP = \frac{13 \times \text{홈런} + 3 \times (\text{볼넷} + \text{몸에 맞는 공}) - 2 \times \text{삼진}}{\text{투구 이닝}} + \text{상수}$$

먼저 FIP는 투수 능력을 측정하는 것이기 때문에 볼넷 계산에서 보통 감독 지시에 따르는 행위인 고의사구는 계산에서 제외한다.¹²⁾

맨 마지막에 나오는 ‘상수’는 리그 FIP를 리그 평균자책점과 똑같이 맞추는 구실을 한다. 따라서 리그 평균자책점에서 ‘ $(13 \times \text{리그 홈런} + 3 \times (\text{리그 볼넷} + \text{리그 몸에 낫는 공}) - 2 \times \text{리그 삼진}) \div \text{리그 이닝}$ ’을 빼면 상수가 나온다. 결국 FIP는 야구팬들에게 친숙한 평균자책점과 비슷한 범위로 투수 능력을 평가할 수 있도록 기능한다.

그래서 FIP를 평균자책점과 비교하는 일도 흔하다. 일반적으로 평균자책점이 FIP보다 크면 해당 투수는 운이 없었다고 평가한다. 이를 다른 말로 표현하면 리그 평균보다 수비 도움을 적게 받았다는 뜻이기도 하다.

이 차이를 토대로 성적을 예측하는 이들도 많다. 불운이 사라지고 나면 성적이 제자리를 찾을 확률이 높기 때문이다. 같은 이유로 팀 동료들 수비 실력이 올라가거나 아니면 팀을 옮겨 더 좋은 야수진을 뒤에 두고 공을 던질 때도 성적 반등을 예상할 수 있다.

반대는 반대다. 거꾸로 평균자책점이 FIP보다 작은 경우에는 행운이 많이 따랐기 때문에 앞으로는 성적이 내려갈 것으로 전망하고는 한다.¹³⁾

사. 리그 이동

리그를 이동하면 선수 ‘능력치’가 바뀐다. BNIP가 능력이라고 해도 모든 상황에서 투수가 같은 수준으로 능력을 유지할 수는 없는 노릇이기 때문이다.

제 아무리 형편없는 프로야구 2군 투수라고 해도 초등학교생들이 뛰는 리틀리그로 간다면 ‘언히터블’ 투수로 변신이 가능할 것이라고 유추할 수 있다. 이런 극단적인 경우가 아니라도 리그 변화에 따라 능력치가 변하리라는 건 상식적으로 예

12) 영어로는 고의사구를 흔히 ‘Intentional Base on Balls(IBB)’라고 표현한다. 자연스레 고의사구를 제외한 볼넷은 ‘Unintentional Base on Balls(UIBB)’가 된다.

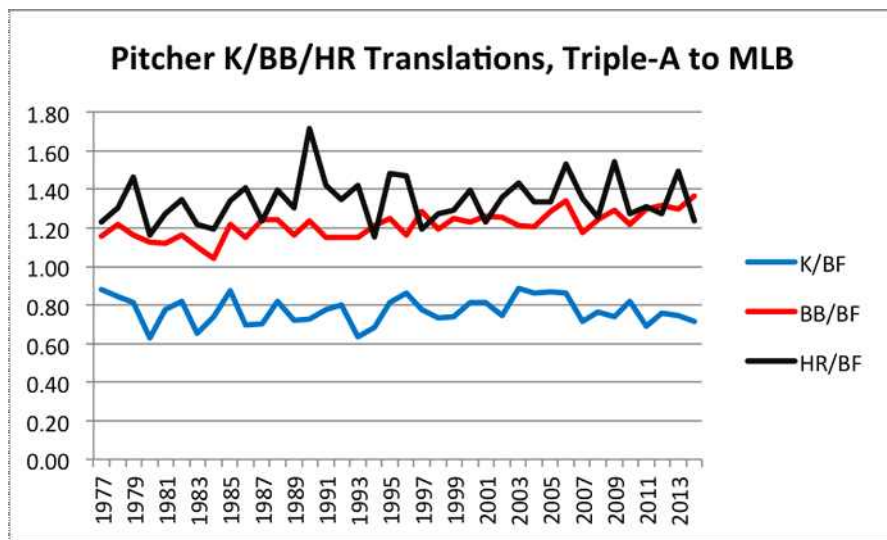
13) www.fangraphs.com/library/pitching/fip

상할 수 있다.

미국 조직 야구에서는 선수가 리그 이동을 하는 일이 많기 때문에 이에 관한 연구가 활발한 편이다. 한 선수가 마이너리그 각 단계를 거치는 것도 리그 이동이고, 메이저리그로 승격하는 것도 마찬가지다. 또 NL과 AL 사이¹⁴⁾에서 선수가 오가는 것 역시 리그 이동이라고 할 수 있다.

리그 이동에 있어 1차적으로 확인할 수 있는 건 리그 수준(환경)에 따라 일정한 방향성을 띤다는 것이다. 이런 방향성은 성적 예측에 도움을 준다. 예를 들어 AAA에서 메이저리그로 승격하는 투수는 전체적으로 투수에게 유리한 삼진 비율은 떨어지고, 투수에게 불리한 볼넷과 홈런 비율은 늘어날 것이라고 예측할 수 있다.

실제로도 이런 변화가 일어난다. [그림 4]는 AAA(Triple-A)에서 뛰던 투수가 메이저리그(MLB)로 올라갈 때 삼진(K) 볼넷(BB) 피홈런(HR) 등 BNIP 비율이 어떻게 변하는지 알려준다.¹⁵⁾



[그림 4] AAA에서 메이저리그 승격 때 BNIP 비율 변화(Lindbergh, 2015)

14) NL에는 지명타자 제도가 없어 투수도 타격을 하기 때문에 리그 득점 환경이 차이가 난다.

15) BF는 Batters Faced를 줄인 말로 '상대 타자 숫자'를 뜻한다.

이런 방향성을 통해서도 BNIP가 능력이라는 사실을 검증할 수 있다. 해외 리그에서 한국 프로야구로 건너오는 외국인 투수를 예로 들자면 이들 가운데 해외 리그에서 삼진 능력이 뛰어났던 투수가 한국 프로야구에서도 계속 뛰어난 삼진 능력을 자랑한다면 삼진이 능력이라고 판단할 수 있는 것이다.

선행 연구 결과는 이런 사실을 지지한다. 이상희(2017)는 2010~2016년 한국 프로야구에서 활약한 외국인 투수 가운데 한국 진출 전 AAA에서 세 시즌 동안 최소 30경기 이상 소화한 선수를 대상으로 BNIP 등에 대한 피어슨 상관관계를 조사했다.

그 결과 99% 신뢰 수준에서 피어슨 상관계수를 따졌을 때 삼진+는 .565를 기록했다. 볼넷+는 이보다 높은 .621다. 한국 프로야구 진출 직전 시즌 10경기 이상 소화한 투수는 역시 같은 신뢰 수준에서 삼진+ .437, 볼넷 + .479였다. 투수가 AAA에서 한국 프로야구로 리그를 바꿀 때도 삼진과 볼넷 능력을 유지한다는 증거다.

아. WAR(Wins Above Replacement)

세이버메트릭스 발전과 함께 각광 받고 있는 또 한 가지 기록이 WAR(Wins Above Replacement)이다. ‘대체 선수 대비 승리 기여도’ 정도로 번역할 수 있는 이 WAR은 해당 선수가 만들어내거나(타자) 막아낸(투수) 점수를 토대로 이 선수가 팀 승수에 얼마나 기여했는지는 알려주는 지표(Schoenfeld, D, 2012)다.

문제는 WAR를 구하는 데 서로 다른 계산법을 활용한다는 점이다. ‘팬그래프스’¹⁶⁾는 투수 WAR 계산에 FIP를 반영해 능력을 평가하려 애쓰고 있는데 이 때문에 오히려 WAR를 제공하는 다른 사이트와 기록 차이가 생긴다.

예컨대 1977~1994년 메이저리그에서 투수로 활약한 잭 모리스(62)의 통산 기록 WAR를 살펴보면 팬그래프스는 55.8¹⁷⁾이라고 계산한 반면 ‘베이스볼 레퍼런

16) 메이저리그 기록 통계 사이트. www.fangraphs.com

17) www.fangraphs.com/statss.aspx?playerid=1009211&position=P

스' 18) 에서는 44.1¹⁹⁾을 제시하고 있다. 만약 55.8을 타율 0.300에 비교할 수 있다면, 43.8은 0.237밖에 되지 않는다.

이게 전부가 아니다. WAR과 사실상 똑같은 개념인 '베이스볼 프로스펙터스' 20) WARP(Wins Above Replacement Player)로 따져 보면 모리스는 통산 64.3²¹⁾승을 기여한 투수로 나온다. 이렇게 차이가 나는 계산 결과를 두고 어느 쪽이 옳다 그르다고 논쟁하는 것 자체가 또 다른 연구 과제가 될 정도다.

게다가 한국 프로야구 WAR을 계산해 제공하는 사이트 역시 서로 다른 계산 방식을 사용한다. 또 이들 사이트에서 활용하는 계산법이 앞서 언급한 메이저리그 기록 사이트에서 사용하는 것과 똑같은 것도 아니다.

이런 상황을 고려하면 한 리그 안에서 같은 계산법을 사용해 분석을 진행하는 게 아주 의미가 없다고는 할 수 없지만, 리그를 바꿔 기록한 WAR이 능력인지 성과인지 따지는 건 큰 의미가 없다고 할 수 있다. 이번 연구에서 WAR를 사용하지 않는 이유다.

3. 투구 추적 시스템

가. 트랙맨(Trackman)

세이버메트릭스는 야구를 과학적으로 연구해 숨은 사실을 밝혀내는 걸 목표로 삼고 있지만 많은 세이버메트리션이 '그런데 너 야구 얼마나 해봤어?' 라는 질문 앞에서는 무기력할 수밖에 없다. 세이버메트리션 대부분이 전문적인 선수 출신이 아니기 때문이다.

특히 이전까지는 세이버메트릭스 연구가 실제 살아 움직이는 투·타구가 아니

18) 메이저리그 및 마이너리그 기록 통계 사이트. www.baseball-reference.com

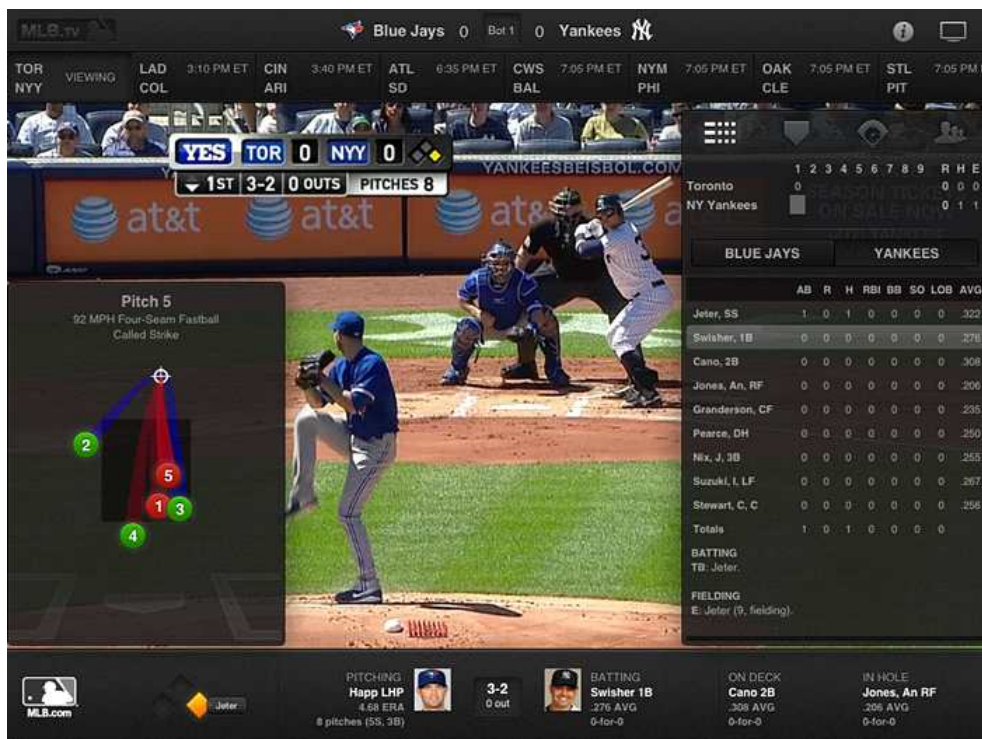
19) www.baseball-reference.com/players/m/morrija02.shtml

20) 세이버메트릭스 커뮤니티. www.baseballprospectus.com

21) legacy.baseballprospectus.com/card/17912/jack-morris

라 이미 플레이가 끝난 상태를 모아 놓은 ‘플레이 바이 플레이(play by play)’ 데이터에 의존하는 일이 많다는 점에서 더욱 그랬다.

세이버메트릭션에게 이런 한계에서 벗어날 수 있도록 도와준 도구가 바로 ‘투구 추적 시스템(PFX·Pitch F/X)’이었다. 스포츠비전이라는 회사에서 제작한 PFX는 투수가 던진 공이 얼마나 빠르지 그 공이 홈플레이트로 날아오는 동안 상하 좌우 얼마나 큰 폭으로 움직이는지도 측정할 수 있다(Silver, N., 2012).



[그림 5] PFX 활용 메이저리그 중계 화면²²⁾

메이저리그 사무국은 2006년 플레이오프부터 PFX를 메이저리그 30개 구장에 모두 설치했다. 그리고 2017 시즌을 시작하며 PFX를 ‘트랙맨(trackman)’²³⁾으로 교

22) ‘At Bat’ 어플리케이션 스크린샷. 왼쪽 아래에 투구 관련 정보가 나타나고 있다.

23) baseball.trackman.com

체했다. PFX가 카메라를 활용한 기술이었다면 트랙맨은 군사용 레이더 기술을 활용한다. 이 때문에 PFX보다 좀더 정교한 측정이 가능하다.

트랙맨은 투수가 던진 공을 타자가 받아쳐 바닥에 떨어질 때까지 Δ 릴리스 포인트²⁴⁾ 위치 Δ 투구 속도 Δ 구종 Δ 상하 변화 Δ 좌우 변화 Δ 타구 속도 Δ 타구 발사 각도 Δ 타구 낙하 위치 등 총 27 가지 정보를 수집한다.



[그림 6] 트랙맨 활용 메이저리그 TV(MLB.tv) 중계 화면²⁵⁾

미국 일리노이대 교수(물리학)이자 세이버메트릭션으로 유명한 앨런 네이션이 “(PFX 등장으로) 우리는 투구 기술에 관해 연구할 수 있는 새롭고 짜릿한 도구를 얻게 됐다(Nathan, 2012)” 고 평가한 건 세이버메트릭스 커뮤니티가 얼마나 이런 도구 등장을 학수고대했는지를 방증한다.

24) 투수가 공을 던질 때 공이 마지막까지 투수 손가락에 붙어 있는 지점

25) 화면 아래를 보면 투수가 던진 공에 관해 초속(Release Velocity), 효과 구속(Effective Velocity·구종과 날아온 위치 등에 따라 타자가 체감하는 속도), 회전수(Spin Rate), 익스텐션(Extension·투수가 공을 던지면서 홈 플레이트 방향으로 내딛은 거리) 등을 측정해 알려준다는 사실을 알 수 있다. 화면 위쪽에 나온 타자 쪽 결과는 타구 속도(Batted Ball Speed), 발사각도(Launch Angle), 비거리(Distance), 비행시간(Hang Time) 등을 표시하고 있다. 땅볼(Ground Ball)이라는 건 공이 곧바로 땅에 맞아 비행시간을 측정하는 게 별 의미가 없다는 뜻이다.

외국인 선수를 뽑을 때 ‘경험’에 의존한다는 건 결국 관찰과 직관에 의존한다는 것과 같은 표현이다. PFX나 트랙맨은 관찰의 폭과 깊이를 모두 더한다. 자연스레 직관을 넘어 데이터에 뿌리를 둔 의사결정을 내릴 수 있도록 도와준다. 21세기는 어떤 투수가 던지는 어떤 공(구종)이 어떻게 좋은지 숫자로 확인할 수 있는 시대다.

나. 구종 가치

구종 가치(Pitch Value) 또는 구종 선형 가중치(Pitch Type Linear Weights)는 이름 그대로 투수가 특정한 구종을 던졌을 때 그 공이 어떤 가치를 가지고 있는지 알려주는 지표다. 많은 세이버메트릭스 지표가 그렇듯이 구종 가치에서 ‘가치’는 점수를 뜻한다.

어떤 투수가 던지는 어떤 공(구종)이 최고인지 따지는 건 사실 야구팬들의 오랜 관심사였다. 예전에는 이를 객관적으로 측정할 수 있는 수단이 없었기에 지루한 논쟁이 이어지기 일쑤였다. 객관적인 측정 도구가 없었기 때문에 객관적인 결론을 내리는 것도 자연스레 불가능에 가까운 일이었다.

PFX 그리고 트랙맨을 통해 투수가 던진 구종을 하나하나 관찰하고 기록할 수 있게 되면서 새로운 가능성이 열렸다. 군사용 레이더는 편견 없이 투수가 던진 공을 측정해 그 결과를 알려준다. 그 결과 야구팬들은 어떤 투수가 던진 어떤 구종이 실제 타자를 상대로 몇 점을 더하고 빼는지 계량적으로 확인할 수 있게 됐다.

‘선형 가중치’라는 표현에서 추론할 수 있듯 구종 가치는 기본적으로 더하기와 빼기를 반복하는 방식으로 계산한다. 어떤 투수가 속구(직구)를 던졌을 때 한 번은 실점을 0.7점 줄이고, 한 번은 0.5점을 늘렸다고 가정해 보자. 그러면 이 투수는 속구 두 개를 던져 $0.2(=0.7-0.5)$ 점을 막아낸 셈이 된다.

특정 기간, 이를 테면 한 시즌 동안 투수가 던진 공에 따라 점수가 올라가고

줄어든 결과를 모두 더해 나오는 숫자가 바로 구종 가치다. 당연히 이 점수가 높을수록 해당 투수가 특정 구종에 강점을 가지고 있다는 뜻이고 낮으면 그 반대로 해석할 수 있다.

그래도 아직 어떤 공이 어떻게 실점을 늘리고 줄이는지에 관한 의문을 풀리지 않았다. 게다가 실제 구종 가치 데이터를 찾아보면 소수점 단위까지 결과가 나오는 걸 알 수 있다. 이런 건 어떻게 계산할 수 있을까.

시작은 볼 카운트를 활용하는 것이다. 프로야구 TV 중계 등에서 스트라이크와 볼 숫자에 따라 ‘투수에게 유리하다’거나 ‘타자에게 유리하다’고 말하는 걸 어렵지 않게 들을 수 있다. 실제로도 볼카운트에 따라 타격 결과가 달라진다.

<표 3> 2012~2014년 한국 프로야구 특정 볼카운트 이후 타격 기록

볼	스트라이크	타율	OPS ²⁶⁾
0	0	.272	.750
	1	.244	.640
	2	.191	.488
1	0	.283	.839
	1	.252	.705
	2	.202	.575
2	0	.296	1.000
	1	.269	.829
	2	.241	.647
3	0	.296	1.249
	1	.289	1.087
	2	.230	0.825

타자에게 볼카운트가 유리할수록 그러니까 스트라이크보다 볼이 많을수록 타자는 더 좋은 타격 기록을 남긴다. 따라서 투수가 특정 구종을 던져 스크라이크를 잡아냈다면 타자의 예상 타격 결과를 끌어내린 셈이 된다. 그리고 그 공은 실점을

26) On base Plus Slugging. 출루율과 장타력을 더한 값으로 타자의 ‘득점 생산력’을 알려주는 세이버메트릭스 지표다.

막아내는 데 도움을 줬다고 할 수 있다.

예를 들어 볼카운트 0볼-0스트라이크에서 커브를 던져 스트라이크를 잡아냈다고 치자. 그러면 이 커브는 원래 .272이던 상대 타자 기대 타율은 .244²⁷⁾로 .028 끌어내렸다. 거꾸로 이 커브가 볼이 됐다면 기대 타율이 .283으로 0.011 올라간다.

이번 설명에서는 타율을 기준으로 삼았지만 타율이 올라갈 때마다 득점이 얼마나 올라가는지 계산하면 이런 변화를 점수 기준으로 바꾸는 것도 가능하다. III-4에서 살펴본 것처럼 각종 기록을 점수 기준으로 바꿔 활용하는 건 세이버메트릭스의 가장 기본적인 접근법이기도 하다.

같은 방식으로 각 타격 결과(단타, 2루타, 3루타, 홈런, 볼넷, 몸에 맞는 공 등)에 따라서도 예상 득점(실점)을 계산할 수 있다.

<표 4> 2017년 한국 프로야구 타격 결과별 득점 가치²⁸⁾

결과	득점 가치	결과	득점 가치
단타	.509	볼넷(UIBB)	.360
2루타	.868	고의사구	.073
3루타	.978	몸에 맞는 공	.432
홈런	1.416	아웃(삼진 제외)	-.287
삼진	-.341		

구종 가치는 기본적으로 BIP 결과를 포함하고 있기 때문에 ‘성과 지표’로 분류할 수 있다. 그러니까 과거에 어떤 구종에 강점을 보였다고 해서 반드시 이듬해에도 그럴 것이라고 예상하는 건 곤란하다. 실제로 연도별 상관계수를 구하면 0.25 미만으로 나타난다.²⁹⁾

그래도 구종 가치가 의미가 있는 건 ‘좋은 공이 좋은 결과를 이끌어낸다’는 인식을 뒤집어 ‘좋은 결과를 낸 공이 좋은 공’이라는 접근법을 우리에게 알려주

27) 0볼-1스트라이크 때 기대 타율

28) 페이스북 그룹 ‘세이버메트릭스 실험실’ bit.ly/2mn1Vkk

29) www.fangraphs.com/library/offense/pitch-type-linear-weights

기 때문이다. 이를 응용해 생각하면 만약 어떤 ‘투수 그룹’ 이 특정 구종에 꾸준히 강점을 보인다면, 그러니까 특정 구종 가치가 높게 나타난다면 해당 그룹은 ‘현재’ 한국 프로야구에서 ‘좋은 공’ 을 던지고 있다고 해석할 수 있다.

모든 투수가 각 구종별을 똑같은 숫자로 던지는 건 아니기 때문에 구종 가치를 비교할 때는 투구 100개 기준으로 환산한 ‘구종 가치/100’ 을 활용하는 게 일반적이다. 따라서 구종 가치/100은 그 공을 100개 던졌을 때 어느 정도로 실점을 줄일거나 늘리는지 알려주는 숫자다.

특정 구종에 대한 리그 전체 평균은 0이며 빠른 공(속구)은 보통 구종 가치가 -1.5~1.5 사이로 나타난다.

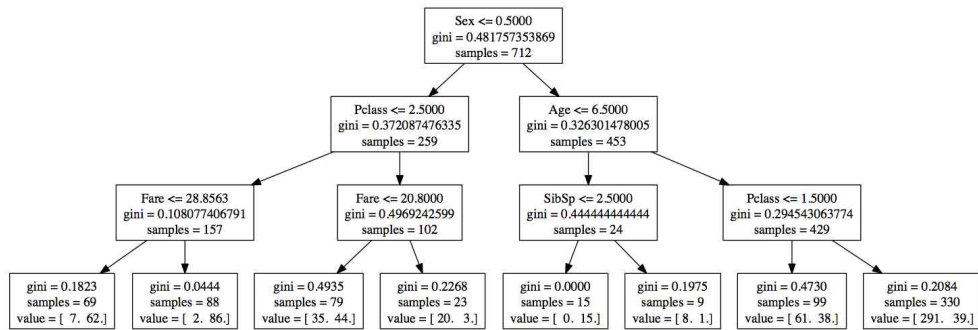
3. 의사결정나무와 확률

가. 의사결정나무

의사결정나무는 의사결정규칙(decision)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다.

분석과정이 나무구조에 의해서 표현되기 때문에 판별 분석(Discriminant Analysis), 회귀분석(Regression Analysis), 신경망(Neural Networks) 등과 같은 방법들에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다(최종후, 서두성, 1999).

의사결정나무 기법은 ‘타이태닉 생존자 예측’ 등 기계 학습(머신 러닝)에서도 널리 쓰인다(Varian, H. R. ,2014).



[그림 7] 의사결정나무 기법 활용 타이태닉 생존자 예측 모델³⁰⁾

관점에 따라서는 외국인 투수 재계약 여부 역시 한국 무대에서 계속 생존할 수 있는지 여부를 묻는 것으로 볼 수 있다.³¹⁾ III-1에서 살펴본 것처럼 한국 프로야구 팀과 재계약한다는 건 연봉이라는 측면에도 선수에게 도움이 되는 행위다. 첫 계약 팀과 실패한 경우에도 팀을 바꿔 한국에 계속 남는 선수들이 있다는 것도 한국이 외국인 투수에게 매력적인 무대라는 증거라고 할 수 있다.

나. 의사결정나무에 대한 수학적 접근

이번 연구에서는 의사결정나무 분석기법 중 CART(Classification And Regression Trees) 알고리즘을 활용했다. CART는 ‘지니 지수(Gini Index)’ 감소량을 이용해 이진분리(Binary Split)를 수행하는 알고리즘이다(Quinlan, 1993). 지니 지수는 불순도(impurity) 또는 다양도(diversity)를 측정한다.

$p(i)$ 를 한 개체가 목표변수의 i 번째 범주에 속할 확률이라고 가정하자. 그러면 목표변수의 i 번째 범주에서 추출한 개체를 j 번째 범주에 속한다고 오분류

30) ericschwarzkopf.de/?p=67

31) 물론 메이저리그 진출을 노리거나 일본 프로야구로 건너가는 등 한국 프로야구보다 상위 리그 팀과 계약하려고 선수가 재계약을 거부하는 사례도 있기 때문에 ‘재계약=생존’ 공식이 100% 맞아 떨어지는 건 아니다.

(misclassification)할 확률은 $p(i)p(j)$ 가 된다.

c 를 목표변수 범주 숫자로 놓고 이렇게 잘못 분류할 확률을 모두 더하는 계산 과정은 아래 식처럼 표현할 수 있다.

$$G = \sum_{j=1}^c \sum_{i \neq j} p(i)p(j)$$

이때 G 는 위와 같은 분류 규칙 아래서 오분류 확률 추정치가 된다.

그 마디에 포함되어 있는 관찰치수를 n , n_i 를 목표변수의 i 번째 범주에 속하는 관찰치 숫자로 놓으면 G 는 아래 식으로 정리 가능하다.

$$G = \sum_{j=1}^c P(j)(1 - P(j)) = 1 - \sum_{j=1}^c P(j)^2 = 1 - \sum_{j=1}^c (n_j/n)^2$$

결국 지니 지수(G)는 원소 n 개 중에서 임의로 2개를 추출했을 때 이 2개가 서로 그룹에 속해 있을 확률을 의미하게 된다. 목표변수 범주가 2개인 경우에 지니 지수는 다음과 같이 표현할 수 있다.

$$G = 2P(1)P(2) = 2\left(\frac{n_1}{n}\right)\left(\frac{n_2}{n}\right)$$

CART 알고리즘은 지니 지수를 가장 감소시켜주는 예측변수와 그 변수의 최적 분리를 자식 마디로 선택한다. n 을 부모마디 관측치 숫자, n_R 과 n_L 을 각각 자식마디 관측치 숫자로 놓으면 지니 지수 감소량은 다음과 같이 계산할 수 있다.

$$\Delta G = G - \frac{n_L}{n} G_L - \frac{n_R}{n} G_R$$

요컨대 CART 알고리즘은 자식 마디로 분리했을 때 불순도가 가장 작도록 가지를 치는 것이다(최종후, 서두성, 1999).

단, 의사결정나무 분석을 진행할 때는 일단 가지를 잘 쳤다고 그 상태로 분석이 모두 끝나는 건 아니다. 의사결정나무 분석을 실시할 때는 먼저 의사결정나무를 형성한 뒤 분류 오류(classification error)를 크게 할 위험이 있거나 부적절한 규칙을 가지고 있는 가지를 제거하는 가지치기(pruning) 작업을 진행하는 게 일반적이다(Berry & Linoff, 1997).

나. 랜덤 포레스트(Random Forest)

랜덤 포레스트는 앙상블(Ensemble) 학습 기법을 활용해 의사결정나무 기법의 단점을 보완하는 기계 학습 기법이다(Breiman, 2001).

앙상블 학습은 데이터로부터 여러 모델을 학습한 다음, 여러 모델의 예측 결과를 종합해 전체 예측 정확도를 높이는 방식이다. 학습 데이터를 무작위로 샘플링해 만든 의사결정나무 여러 개를 사용하기 때문에 랜덤 포레스트라는 이름이 붙었다(クジラ飛行机, 2016).

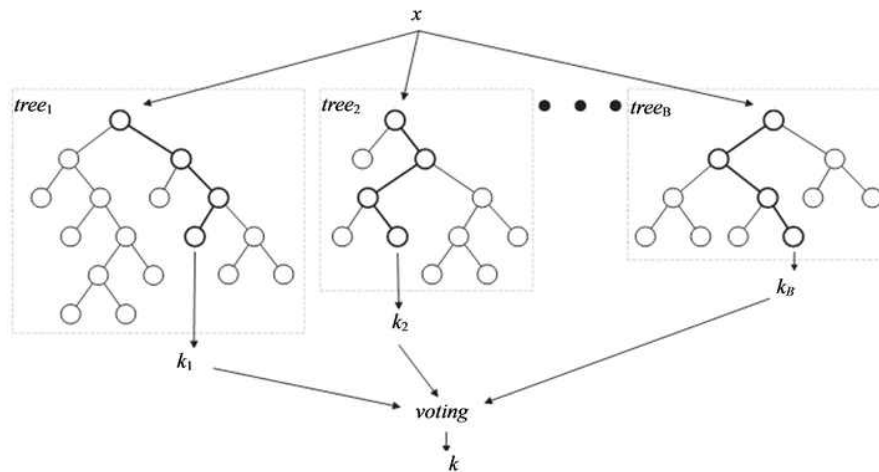
결국 의사결정나무 모형을 기저로 무작위성을 최대로 부여함으로써 예측오차를 줄이는 기법이 바로 랜덤 포레스트다(윤진은, 2015).

랜덤 포레스트는 두 가지 방법을 사용해 다양한 의사결정 나무를 만든다. 첫 번째는 의사 결정 나무를 만들 때 데이터의 일부를 복원 추출로 꺼내고 해당 데이터에 대해서만 의사결정나무를 만드는 방식이다. 즉, 각 데이터의 일부만 사용해 의사결정 나무를 만든다. 두 번째는 노드(node) 내 데이터를 자식 노드로 나누는 기준을 정할 때 전체 변수가 아니라 일부 변수만 대상으로 해 가지를 나눌 기준을 찾는 방법이다.

새로운 데이터에 대해 예측을 수행할 때는 의사결정나무 여러 개가 내놓은 예측 결과를 투표(voting) 방식으로 합한다. 예를 들어 의사결정나무 5개 가운데 Y를

예측한 나무가 3개, N을 예측한 나무가 2개면 Y를 최종 결과로 결정하는 방식이다.

랜덤 포레스트는 일반적으로 성능이 뛰어나고 의사결정나무 하나가 아니라 여러 개를 사용해 과(過)적합 문제를 피한다(서민구, 2014)



[그림 8] 랜덤 포레스트 구조(Verikas, Gelzinis, & Bacauskiene, 2011)

IV. 연구 방법

1. 연구 대상

2007년 이후 2016년까지 한국 프로야구 구단과 계약한 외국인 투수 중 1군 기록이 있는 130명을 연구 대상으로 삼았다.

원래 이 기간 한국 프로야구 무대에 첫 선을 보인 선수는 총 139명이다. 이 중 랜들(40), 레스(44), 리오스(45 · 이상 두산), 브라운(40 · 삼성, LG), 왈론드(41 · 두산), 존슨(42 · SK), 카브레라(45 · 롯데), 켈러웨이(42 · 현대), 하리칼라(46 · LG) 등 9명은

2006년 이전에 한국 팀과 첫 계약을 맺은 적이 있어 연구 대상에서 제외했다.

외국인 투수가 재계약에 성공해 누적 기록이 있는 경우에도 한국 프로야구 성적은 첫 해 기록만 연구 대상으로 활용했다. 첫해 재계약을 결정할 때는 미래 성적을 모르는 상태에서 재계약을 결정하기 때문이다.

해외 성적 관련 연구에서는 다른 수준에 속한 조직 야구 팀, 미국 독립리그 소속팀, 일본 및 대만 프로야구 등에서 뛰다 한국 프로야구 팀과 계약한 경우에는 자료 수집 대상에서 제외했다. 한국 프로야구 데뷔 직전에 AAA에서 뛴 외국인 투수는 총 112명(86.2%)이었다.

2. 연구 절차

가. 한국 프로야구 첫 시즌 기록 수집 및 정리

외국인 선수가 한국 프로야구 무대에서 남긴 기록은 KBO 홈페이지에 올라 있는 ‘2015 KBO 기록대백과 제5판’ 과 2015 ~ 2017년 ‘프로야구 연감’ 을 통해 확보했다.

2003년 초판이 나온 기록대백과는 한국 프로야구에서 (투수 기준으로) 공을 한 개라도 던진 선수 기록을 1100여 페이지에 걸쳐 모두 정리한 책이다. KBO에서 해마다 펴내는 프로야구 연감은 전년도 프로야구를 결산하는 책으로 선수 개인 성적을 포함하고 있다.

한국 프로야구 기록을 제공하는 인터넷 사이트가 없는 건 아니지만 기록이 부정확한 경우가 없지는 않다. 다소 불편하더라도 KBO에서 제공하는 기록을 직접 활용하기로 한 이유다.

KBO는 이들 자료를 모두 PDF 파일 형태로 올려두고 있다. 이 파일을 내려 받아 연구 대상에 포함된 각 선수 항목을 찾아 필요한 기록을 정리했다. 그리고 이를 마이크로소프트(MS) <표 5> 같은 구조로 엑셀 파일에 정리해 연구해 활용했다.

<표 5> 외국인 투수 데이터 파일 구조

열	항목	열	항목	열	항목
A	이름	K	9이닝당 삼진	W	승
B	출생연도	L	9이닝당 볼넷	X	패
C	키	M	삼진 대 볼넷 비율	Y	세이브
D	몸무게	N	9이닝당 피홈런	Z	홀드
E	bbref 링크	O	FIP	AA	이닝
F	재계약 여부	P	BABIP	AB	실점
G	총시즌	Q	경기당 평균 이닝	AC	자책점
H	연도	T	경기	AD	상대타자
I	팀	U	선발	AE	피안타
J	나이	V	평균자책점		

E열에 들어 있는 ‘bbref 링크’는 앞에서 III-10에서도 등장한 적 있는 ‘베이스볼 레퍼런스’ 선수 페이지 인터넷 주소(URL)를 가리킨다. 이 URL은 해외 리그 기 수집이 1차 목적이며, 파일과 파일 사이에 선수를 연결하는 코드 구실도 맡았다.

F열 재계약 여부는 n년도에 계약한 팀과 n+1년도에도 계약했을 때만 재계약을 했다고 봤다(재계약 1, 미계약0). 나이는 ‘해당연도 - 출생연도’를 통해 계산했다.

삼진, 볼넷, 피홈런 등 BIP를 9이닝을 기준으로 환산한 건 선수 마다 등판 경기 숫자 및 투구 이닝이 다르기 때문에 이를 비율로 환산한 것이다. 야구에서는 이런 차이를 감안해 투수 성적을 평가할 때 9이닝당 기록을 널리 쓰고 있다. 투수 성적을 나타내는 가장 대표적인 지표인 평균자책점부터 ‘자책점 $\times$ 9 $\div$ 투구이닝’으로 계산하는 ‘9이닝당 자책점’이다.

삼진 대 볼넷 비율 역시 세이버메트릭스에서 투수 제구력을 측정할 때 널리 쓰이는 지표다. 삼진 대 볼넷 비율은 삼진이나 볼넷만으로는 확인할 수 없는 정보를 우리에게 알려준다.

소위 ‘파워(power) 피쳐’ 중에는 삼진을 많이 잡지만 그만큼 볼넷도 많이 내줘 그다지 성공적이지 않은 투수가 많다. 거꾸로 ‘제구형(finesse) 투수’ 중에는 삼진은

적어도 볼넷도 적어 성공을 거둔 사례도 많다.³²⁾ 따라서 이미 삼진과 볼넷을 수집했지만 이와 별개로 이 비율을 따로 정리했다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	이름	출생연도	키	몸무게	BBREF	재계약	종시준	연도	팀	나이	K	BB	KBB	HI
2	레이번	1974	193	95	http://www	1	2	2007	SK	33	4.78	4.19	1.14	
3	로드리게스	1972	185	77	http://www	0	1	2007	KIA	35	4.26	5.68	0.75	
4	로마노	1972	188	88	http://www	0	1	2007	SK	35	4.98	4.12	1.21	
5	리오스	1972	188	86	http://www	0	1	2007	두산	35	5.64	2.22	2.53	
6	매존	1976	188	99	http://www	0	1	2007	삼성	31	5.47	3.24	1.69	
7	세드릭	1978	188	99	http://www	0	1	2007	한화	29	7.96	5.91	1.35	
8	스코비	1979	185	88	http://www	0	1	2007	KIA	28	4.13	4.42	0.94	
9	에서튼	1976	185	88	http://www	0	1	2007	KIA	31	6.75	3.38	2.00	
10	옥스포링	1977	185	88	http://www	1	2	2007	LG	30	4.57	3.57	1.28	
11	윌슨	1976	193	99	http://www	0	1	2007	삼성	31	2.13	2.61	0.82	
12	하리칼라	1971	188	83	http://www	0	1	2007	LG	36	2.50	3.96	0.63	
13	다카쓰	1968	183	81	http://www	0	1	2008	넥센	40	7.71	2.14	3.60	
14	데이비스	1975	190	87	http://www	0	1	2008	KIA	33	6.20	4.57	1.35	
15	디아즈	1980	185	81	http://www	0	1	2008	KIA	28	4.57	3.77	1.21	
16	레스	1973	188	83	http://www	0	1	2008	두산	35	5.97	3.69	1.62	
17	레이	1974	188	90	http://www	0	1	2008	SK	34	4.87	6.64	0.73	
18	레이어	1977	188	92	http://www	0	1	2008	두산	31	5.14	2.57	2.00	
19	리마	1972	188	77	http://www	0	1	2008	KIA	36	3.60	3.99	0.90	
20	매클레리	1974	196	99	http://www	0	1	2008	롯데	34	3.81	4.60	0.83	
21	브라운	1977	188	90	http://www	0	2	2008	LG	31	3.72	3.22	1.15	
22	선	1977	193	104	http://www	0	1	2008	삼성	31	4.15	5.54	0.75	
23	얀	1975	193	115	http://www	0	1	2008	SK	33	5.22	2.45	2.13	
24	에니스	1979	196	99	http://www	0	1	2008	삼성	29	6.07	2.12	2.86	
25	오버몰러	1976	188	88	http://www	0	1	2008	삼성	32	5.38	3.18	1.69	
26	코르테스	1973	180	88	http://www	0	1	2008	롯데	35	7.11	2.13	3.33	
27	쿠비안	1972	188	74	http://www	0	1	2008	SK	36	3.86	10.29	0.38	
28	토마스	1977	193	106	http://www	1	2	2008	한화	31	8.95	3.27	2.74	
29	구롭슨	1977	188	83	http://www	0	1	2009	KIA	32	5.30	2.57	2.07	
30	글로버	1976	196	90	http://www	1	3	2009	SK	33	7.67	2.21	3.46	
31	나이트	1975	183	79	http://www	1	2	2009	삼성	34	7.57	4.15	1.82	
32	니코스키	1973	188	83	http://www	0	1	2009	두산	36	6.13	4.30	1.42	

[그림 9] 외국인 투수 파일 스크린샷

나. 마이너리그 AAA 기록 수집 및 정리

마이너리그 기록은 bbref 자료를 수집해 활용했다. 컴퓨터 프로그래밍 언어 R에서 ‘rvest’³³⁾ 패키지를 활용해 bbref 사이트 내용을 긁어오는 크롤러(crawler)를 만들어 외국인 투수 데이터를 확보했다. BNP를 중심으로 마이너리그 자료를 정리

32) m.mlb.com/glossary/advanced-stats/strikeout-to-walk-ratio. 흔히 파워 피쳐와 제구형 투수는 ‘(삼진 + 볼넷)÷투구 이닝’으로 계산하는 ‘PFR(Power Fineess Ratio)’에 따라 평가되지만 명확히 합의한 기준이 있는 건 아니다. 본 연구에서는 이 두 그룹을 꼭 나눌 필요가 있는 건 아니기 때문에 이에 대한 추가 설명은 생략한다.

33) cran.r-project.org/web/packages/rvest/rvest.pdf

한 MS 엑셀 파일 구조는 <표 6>과 같다.

<표 6> 마이너리그 기록 데이터 파일 구조

열	항목	열	항목	열	항목
A	이름	E	리그 레벨	I	볼넷
B	연도	F	이닝	J	몸에 맞는 공
C	팀	G	상대 타자	K	삼진
D	리그명	H	피홈런		

League-by-League Pitching

Share & more ▼ Glossary

Rk	Lg	Lev	Tms	PAGE	R/G	W	L	W-L%	ERA	G	GS	GF	CG	SHO	SV	IP	H	R	ER	HR	BB
1	American League	Major League	15	28.6	4.47	1229	1199	.506	4.20	2428	2428	2384	44	122	644	21611.0	21293	10856	10089	2934	7280
2	National League	Major League	15	28.1	4.48	1198	1228	.494	4.16	2428	2428	2389	39	154	632	21695.1	20983	10888	10035	2676	7808
3	International League	AAA	14	26.8	3.96	1005	1005	.500	3.64	2010	2010	1955	55	193	504	17673.2	16999	7952	7149	1418	5991
4	Mexican League	AAA	16	29.1	4.71	888	888	.500	4.34	1782	1782	1736	46	118	427	15544.2	16772	8401	7489	1277	5751
5	Pacific Coast League	AAA	16	26.9	4.78	1137	1137	.500	4.45	2278	2278	2240	38	130	558	19927.0	20729	10890	9862	1932	7391
6	Eastern League	AA	12	24.5	4.34	846	846	.500	3.95	1692	1692	1644	48	103	424	14792.1	14594	7351	6494	1201	5447
7	Southern League	AA	10	24.3	4.08	695	695	.500	3.69	1390	1390	1366	24	106	378	12221.1	11474	5669	5014	808	4487
8	Texas League	AA	8	24.4	4.12	559	559	.500	3.69	1118	1118	1104	14	91	300	9887.1	9313	4610	4059	871	3360
9	California League	Adv A	10	23.0	5.01	700	700	.500	4.35	1400	1400	1388	12	59	331	12448.1	12645	7018	6023	1197	4573
10	Carolina League	Adv A	8	22.9	4.52	556	556	.500	4.07	1112	1112	1089	23	70	290	9634.0	9510	5023	4352	677	3833
11	Florida State League	Adv A	12	23.1	4.04	815	815	.500	3.58	1632	1632	1587	45	144	416	14222.0	13450	6597	5657	870	5138
12	Midwest League	A	16	21.7	4.05	1114	1114	.500	3.51	2230	2230	2199	31	179	591	19591.2	18417	9033	7649	1062	6773
13	South Atlantic League	A	14	21.8	4.20	975	975	.500	3.68	1950	1950	1920	30	132	525	17027.1	16111	8193	6953	1218	5769
14	New York-Pennsylvania League	Short-Season A	14	21.3	4.03	528	528	.500	3.41	1056	1056	1050	6	84	268	9246.1	8438	4254	3501	398	3400
15	Northwest League	Short-Season A	8	21.2	4.82	301	301	.500	4.04	602	602	602	0	39	150	5308.1	5185	2899	2382	279	2281
16	Appalachian League	Rookie	10	20.9	4.85	337	337	.500	4.10	674	674	662	12	45	157	5812.1	5723	3270	2647	440	2154
17	Arizona League	Rookie	14	20.9	5.00	386	386	.500	3.97	772	772	772	0	42	198	6787.1	6732	3863	2993	271	2504
18	Gulf Coast League	Rookie	17	20.7	4.12	481	481	.500	3.40	966	966	964	2	88	235	8361.0	7577	3983	3161	339	2982
19	Pioneer League	Rookie	8	21.4	5.91	301	301	.500	4.99	602	602	594	8	12	122	5279.1	6019	3557	2927	490	1921
20	Australian Baseball League	Foreign	6	25.9	4.78	119	119	.500	4.28	238	238	229	9	12	50	2016.1	2053	1138	958	171	812
21	Cuban National Series	Foreign	16	28.9																	
22	Dutch Major League	Foreign	8	28.5	5.09	167	167	.500	4.18	336	336	318	18	36	61	2875.2	2875	1711	1337	71	1536
23	Italian Baseball League	Foreign	7	28.9	4.07	115	115	.500	3.42	234	234	228	6	30	43	2016.0	1792	953	767	46	938
24	Japan Central League	Foreign	6	28.0	4.17	415	428	.492	3.69	826	826	810	47	22	200	7657.0	7389	3441	3143	725	2675
25	Japan Pacific League	Foreign	6	27.5	4.05	425	412	.508	3.66	849	849	799	46	17	219	7683.0	7409	3435	3121	616	2781
26	Korean Baseball Organization	Foreign	10	28.5	5.73	719	725	.498	5.20	1431	1431	1437	18	7	325	12980.2	14817	8203	7506	1511	5432

[그림 10] bbref 리그 페이지 스크린샷

이 연구에서는 한국 프로야구 진출 1년 전 AAA 기록만 활용했다. 원래 Marcel 등 야구 선수 성적 예측 시스템(projection)을 활용할 때는 3년치 데이터를 수집하는 게 일반적이다(Tango, T. M., Lichtman, M. G., & Dolphin, A. E., 2007).

이는 선수가 갑자기 성적이 좋아졌을 때 이 해가 우연히 운이 좋은 ‘플루크

(fluke) 시즌’ 인지 아니면 알을 깨고 나온 ‘브레이킹(breaking) 시즌’ 인지 알기가 어렵기 때문이다.

하지만 한국 프로야구에서 외국인 투수를 영입하는 과정은 메이저리그에서 선수 성장을 기다리는 것과 사정이 다르다. 메이저리그 팀은 플루크라고 판단한 선수를 좀 더 지켜볼 여유가 있지만, 한국 구단에서 볼 때는 이 선수가 메이저리그행 가능성이 있다는 건 영입 경쟁에서 밀릴 수밖에 없는 이유가 된다.

거꾸로 꾸준히 좋은 성적을 내다가 갑자기 슬럼프에 빠진 선수는 한국 프로야구 구단에서 영입할 가능성이 높은 후보군에 이름을 올릴 확률이 높다. 이 역시 가장 최근 1년 기록을 연구 대상으로 삼은 이유다.

다. 리그 평균 기록 수집 및 정리

Ⅲ-1에서 살펴본 것처럼 야구에서는 리그마다 경기당 평균 득점 환경을 통해 살펴 볼 수 있는 득점 환경이 다르다(공시적 차이). 또 같은 리그라고 해도 득점 환경이 다를 수 있다(통시적 차이).

이런 차이는 선수를 평가할 때 기준점이 서로 달라져야 한다는 뜻이기도 하다. 이럴 때 사용하는 도구가 바로 ‘기록+’ 다. 기록+ 계산에는 리그 (평균) 기록이 필요하다. 이번 연구에서는 한국 프로야구와 AAA 소속 3개 리그를 연구 대상으로 삼고 있기 때문에 이 3개 리그의 연도별 기록이 필요하다고 할 수 있다.

<표 7> 리그 기록 데이터 파일 구조

열	항목	열	항목	열	항목
A	연도	E	이닝	I	몸에 맞는 공
B	리그명	F	전체 타자(타석)	J	삼진
C	경기당 평균 득점	G	홈런		
D	평균자책점	H	볼넷		

문제는 스탯티즈에서 2014년 이후 구종 가치만 제공하고 있다는 점이다. 그 전에는 한국 프로야구 각 구장에 관련 측정 장비가 들어오지 않았던 상태라 관련 자료를 확보하기가 어려웠다. 이런 이유로 ‘구종 가치’ 관련 연구에서는 표본 확보 차원에서 누적 기록 전체를 활용했다.

스탯티즈에서 2014~2016년 구종 가치를 확인할 수 있는 외국인 투수는 중복 그러니까 한 투수의 여러 해 기록을 별개 선수로 구분하면 총 76명이다.

이 중에서 시스코(34 · 전 kt) 유먼(38 · 전 롯데, 한화)은 한국 프로야구 팀과 계약하기 전 미국 독립리그에서, 홀튼(38 · 전 KIA)은 일본 프로야구에서 뛰었기 때문에 연구 대상에서 제외된 상태라 구종 가치 자료에서도 제외했다.

또 유먼은 2014년에는 롯데, 2015년에는 한화에서 뛰었기 때문에 결국 연구 대상에 포함된 외국인 투수는 총 72명이다.

마. 기록 분석

기본적인 통계 처리에는 ‘MS 윈도 10 64비트’ 환경에서 MS 엑셀과 함께 통계 프로그램 ‘R x86 3.4.2’ 를 활용했다.

‘의사결정나무’ 분석을 실시할 때는 ‘rpart³⁵⁾’ 를 사용했다. rpart는 R에서 CART 알고리즘을 활용해 의사결정나무 분석을 실시할 수 있도록 도와주는 패키지다.

‘랜덤포레스트’ 분석에는 ‘randomForest³⁶⁾’ 패키지를 이용했다. randomForest는 이름 그대로 R에서 랜덤포레스트 분석을 가능하게 만들어준다.

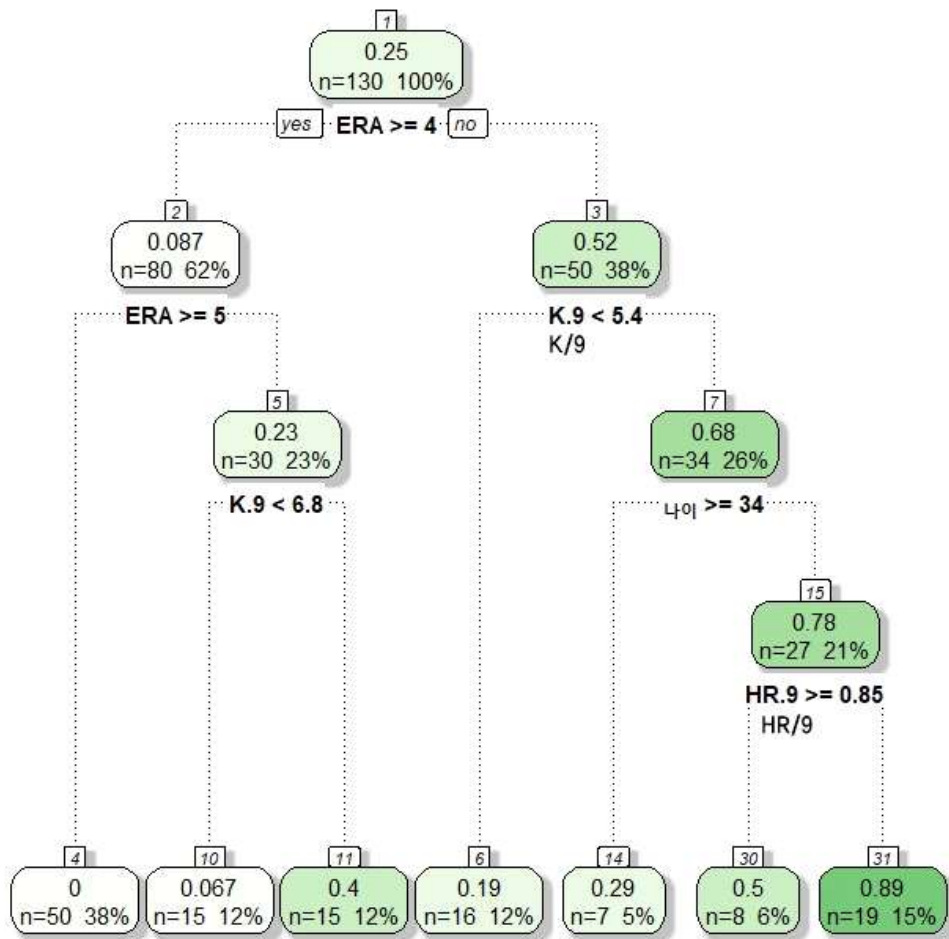
35) cran.r-project.org/web/packages/rpart/rpart.pdf

36) cran.r-project.org/web/packages/randomForest/randomForest.pdf

V. 연구 결과

1. 외국인 투수 재계약 예측 의사결정나무 모델

중속변수(목표변수)에 재계약(0 미계약, 1 재계약), 독립변수(입력변수)에 키, 몸무게, 나이, 9이닝당 삼진(K), 9이닝당 볼넷(BB), 9이닝당 피홈런(HR), BABIP, 평균자책점(ERA)을 투입해 의사결정나무 분석을 실시하면 [그림 10]과 같은 결과가 나온다.

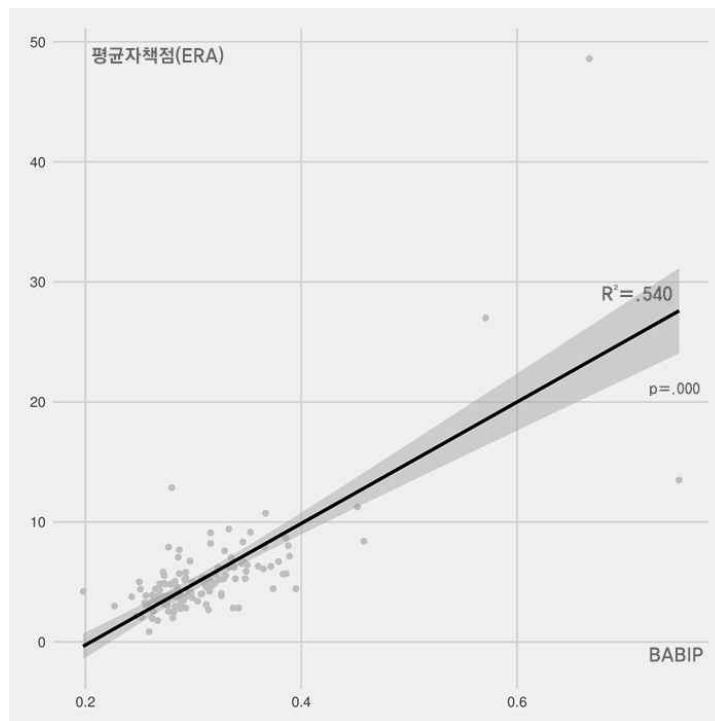


[그림 12] 외국인 투수 재계약 여부 의사결정나무 모델³⁷⁾

이를 통해 평균자책점이 외국인 투수 재계약에 제일 큰 영향을 끼치는 변수라는 사실을 확인할 수 있다. 평균자책점 4 미만일 때는 재계약 성공률이 49.1%(53명 중 26명)로 오른다. 또 재계약에 성공한 33명 중 76.5%(26명)가 평균자책점 4 미만이었다.

문제는 평균자책점이 ‘능력’이 아니라 ‘성과’라는 점이다. 당연히 해마다 변화가 크다. J. C. Bradbury(2015)에 따르면 n 년과 $n+1$ 년 평균자책점 사이 상관계수(R^2)는 0.13밖에 되지 않는다.

외국인 투수도 이런 변화를 경험한다. 한국 프로야구 데뷔 첫 해 재계약에 성공한 선수들 기록을 한 데 모아 평균자책점을 계산하면 3.60이 나오지만 그 이듬해에는 4.20으로 오른다.



[그림 13] 외국인 투수 평균자책점과 BABIP 상관관계

37) R에서는 변수 이름에 '/' 기호를 사용할 수 없어 '.'으로 나왔다.

이는 평균자책점이 BABIP에 영향을 많이 받는 기록이기 때문이다. BABIP를 독립변수, 평균자책점을 종속변수로 놓고 선형회귀분석을 실시하면([그림 13]) 신뢰 수준 99%($p < .000$)에서 R^2 .540가 나온다.

III에서 살펴본 것처럼 BABIP를 결정하는 가장 큰 요소는 (행)운이다. 제아무리 삼진을 많이 잡고, 볼넷을 적게 내주고, 홈런을 맞지 않는 투수라고 해도, 연거푸 바가지 안타를 맞으면 점수를 내주게 마련이다.

이런 실점이 쌓이면 결국 평균자책점이 올라가게 된다. 그런데도 각 구단에서는 평균자책점을 기준으로 외국인 투수를 걸러내고 있다. 이번 해에 (행)운이 따른 선수가 다음 해에도 행운이 따를 것이라고 기대하며 재계약을 선물하는 것이다.

2. BNIP 대(對) BIP

그렇다고 재계약에 성공한 투수들이 능력에 해당하는 BNIP가 떨어지는 건 아니다. 재계약 여부를 ‘집단변수’로 놓고 9이닝당 삼진(K), 9이닝당 볼넷(BB), 9이닝당 피홈런(HR)에 대해 독립표본 T검정을 실시하면 재계약에 성공한 선수가 신뢰도 95% 수준에서 이 세가 기록 모두 뛰어난 결과를 거둔 것으로 나타났다³⁸⁾

<표 8> 재계약 여부에 따른 외국인 투수 BNIP 기록 비교

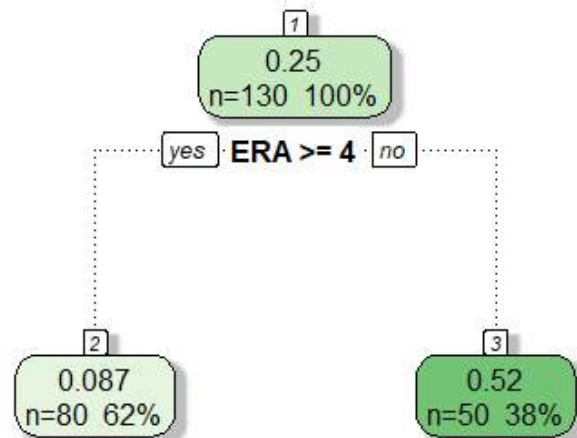
기록	재계약	N	평균	표준편차	t값	p값
K	X	97	6.22	1.81	-2.47	.017
	O	33	7.13	1.84		
BB	X	97	4.44	2.60	2.64	.009
	O	33	3.21	1.06		
HR	X	97	.92	0.68	2.23	.024
	O	33	.63	0.31		

38) 같은 리그 역시 시대에 따라 득점환경이 달라지기 때문에 기록+를 쓰는 게 맞지만 논의 전개상 각 프로야구단에서는 겉으로 드러난 성적을 통해 재계약 여부를 확인한다는 차원에서 일단 기록을 직접 활용했다.

자연스레 K, HBB, HR을 계산에 사용하는 FIP 역시 99% 신뢰 수준($p < .000$)에서 재계약에 성공한 선수는 3.89, 실패한 선수는 4.97로 나타난다. FIP는 정의 그대로 순수한 투수 능력을 보여주는 지표다.

그런데 V-1에서 살펴본 의사결정나무 모델(1)에 FIP를 독립 변수로 추가 투입해도 이 모델(1-1)은 아무 변화도 일어나지 않는다. FIP가 외국인 선수 재계약에 별다른 영향을 주지 못하고 있는 것이다. 이를 다른 말로 표현하면 투수 능력이 재계약에 영향을 끼치지 못하고 있다고 할 수 있다.

오히려 이번에는 교차 검증(CV · Cross Validation)을 통해 가지치기 작업을 진행하면([그림 12]) 평균자책점이 재계약을 결정하는 사실상 유일한 평가 기준으로 바뀐다. FIP를 투입하면 평균자책점 중요도만 올라가는 셈이다.



[그림 14] 외국인 선수 재계약 여부 의사결정나무 모델 가지치기

투수가 자기 실력을 충분히 발휘하고 있는 데도 여러 불운이 겹치면 겉으로 드러난 성과에 문제가 있는 것처럼 보일 수밖에 없다. 물론 이렇게 불운에 시달린 투수를 꼭 구제해야 한다는 법이 없는 것처럼 이런 선수들이 앞으로도 계속 불행에 시달릴 것이라고 전망할 필요도 없다.

거꾸로 운이 좋았던 선수가 계속 운이 좋을 것이라고 전망하는 건 어떨까. 일단

최근 10년간 한국 프로야구 구단은 행운이 함께한 외국인 선수를 포기하지 않았다. 이번 연구에서 행운을 대표하는 기록인 BABIP 역시 99% 신뢰 수준($p < .000$)에서 재계약에 성공한 선수가 0.282로 재계약하지 못한 선수(0.316)보다 더 좋은 결과를 얻었다.

성공하려면 물론 실력(능력)만큼 행운도 필요하다. 하지만 행운을 실력이라고 착각하는 건 결국 모든 걸 행운에 맡기겠다는 생각과 차이가 없다. 혹시 행운이 쌓이면 실력이 되지는 않을까.

3. 삼진, BABIP 그리고 재계약

야구에서 투수에게 제일 중요한 임무는 아웃 카운트를 잡아내는 것이고, 투수가 아웃 카운트를 잡아내는 길은 딱 두 가지뿐이다. 직접 삼진을 잡거나 아니면 맞춰 잡거나. 9이닝당 탈삼진(K)은 삼진 비율을, BABIP는 맞춰 잡은 비율을 보여준다.

[그림 15]는 연구 대상에 포함된 투수 기록을 K(x축)와 BABIP(y축)에 따라 그린 그래프다. 그래프 가운데 짙은 선은 2007~2016년 평균치(BABIP 0.315, K 6.60)다.

재계약에 성공한 외국인 투수(주황색 점) 중에서 한국 프로야구 데뷔 첫 시즌 BABIP가 0.318을 넘어가는 선수는 단 한명도 없다.

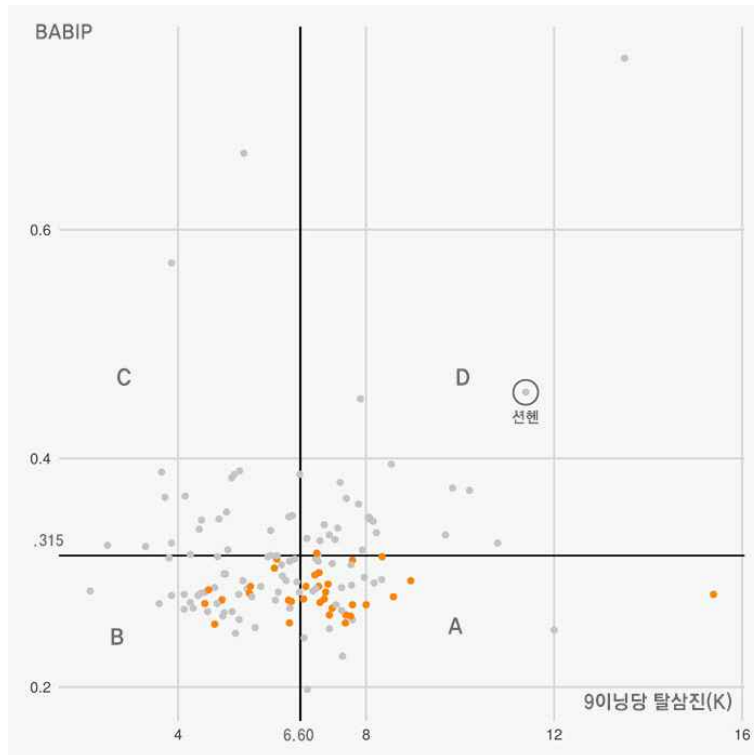
K는 ‘성과인 동시에 능력’이지만 BABIP는 아니다. 그런데도 현재 프로야구 팀에서는 BABIP 그리고 BABIP에 영향을 받아 역시 능력보다 성과에 가까운 평균자책점을 토대로 외국인 투수 재계약 여부를 결정하고 있는 것이다.

이런 이유로 2009년 한화에서 썬 선헨(36)처럼 삼진을 리그 평균보다 172% 많은 9이닝당 11.4개를 잡은 투수도 BABIP가 0.458이나 되면 시즌 중반에 짐을 싸서 돌아가야 했다. 선헨은 이듬해 AAA 라스베이거스에서 57과 3분의 2이닝을 평균자책점 2.81로 막아냈고, 뉴욕 메츠 유니폼을 입고 메이저리그 무대에 서기도 했다.

물론 선헨은 극단적인 케이스지만³⁹⁾ BABIP 그리고 이에 영향을 많이 받는 평

39) 선헨은 한국 프로야구에서 평균자책점 8.40을 기록했는데 당시 FIP는 3.50밖에 되지 않았다.

균자책점 때문에 국내 프로야구 구단이 실제로 성적이 반등할 수 있는 선수를 내보내고, 엉뚱한 투수를 찾는 데 시간과 에너지를 낭비하고 있을 수 있다는 개연성은 확인할 수 있다. 참고로 2017년 9월까지 한화에서 뛰다 재계약에 실패한 외국인 투수 중 8명이 메이저리그 무대를 밟았다. 한화 한 팀만 따져도 무시하기 힘든 숫자다.



[그림 15] 외국인 투수 K-BABIP 분류

4. 투수 유형별 마이너리그 복귀 성적 비교

그러면 마이너리그 AAA로 돌아간 투수들은 어떤 성적을 남겼을까. 이 결과를 알아보면 한국 프로야구 팀이 ‘행운’을 기준으로 외국인 선수 재계약 여부를 판

단한 게 옳은 일이었는지를 확인할 수 있다.

이를 알아보려면 먼저 투수 유형을 구분하는 작업이 필요하다. [그림 15]에서 A, B, C, D는 K와 BABIP 성과에 따른 각 투수별 유형을 나타낸다.

<표 9> 투수 유형별 특징

유형	인원	K	BABIP	평균자책점 ⁴⁰⁾
A	39	평균 이상	평균 이상	3.78
B	50	평균 이하	평균 이상	4.09
C	19	평균 이하	평균 이하	6.80
D	22	평균 이상	평균 이하	5.72

외국인 투수가 한국 프로야구에서 거둔 성적이 행운에 영향을 많이 받은 결과라면 AAA 복귀 후 D형은 성적이 좋아지고, B형은 성적이 나빠질 것으로 예상할 수 있다. D형은 삼진 능력이 있지만 BABIP 도움을 받지 못한 유형이고 B형은 반대로 운이 많이 따른 케이스이기 때문이다

따라서 이번에는 능력이 아니라 성과를 측정해야 한다. 이 연구에서 성과를 확인할 수 있는 대표 지표는 평균자책점이다. 그리고 9이닝당 탈삼진(K) 역시 투수 능력인 동시에 아웃 카운트를 잡아 낸 성과다.

단, 이 외국인 투수들이 AAA에 복귀한 시기와 리그가 모두 다르기 때문에 추가 연구가 필요하다. 이럴 때 도움을 받을 수 있는 게 바로 ‘기록+’다. AAA 복귀 첫째 성적을 토대로 개인별 ‘평균자책점+’를 계산한 뒤 선수 유형에 따라 일원배치분산분석(ANOVA)을 실시했다.⁴¹⁾

평균자책점은 낮을수록 좋은 기록이기 때문에 평균자책점+ 역시 낮을수록 좋은 결과다. 통계적으로 90% 신뢰 수준을 넘어서는 결과($p=.133$)지만 이번 연구에서 가정했던 내용이 아주 엉뚱한 이야기만은 아니었다는 사실을 확인할 수 있다.⁴²⁾

40) 각 유형에 해당하는 투수들 전체 기록을 가지고 계산한 값

41) 여기서 평균이라고 나오는 자료는 선수 전체 기록을 더해 ‘평균자책점+’를 계산한 자료가 아니라 이 선수들 평균자책점+의 평균이다.

42) 평균자책점+ 역시 성과를 측정한다는 점을 명심해야 한다. 그러니까 행운이 큰 영향을 끼친다.

<표 10> 투수 유형별 AAA 복귀 후 평균자책점+

유형	인원	평균	표준편차	비고
A	11	106.3	18.1	F 1.906 p .133
B	26	235.6	657.4	
C	16	99.8	38.9	
D	13	93.4	24.0	

삼진 능력은 있지만 BABIP에서 도움을 받지 못했던 D형은 AAA로 돌아가면 성과가 좋아진다. 거꾸로 삼진 능력은 없지만 BABIP 도움으로(행운이 따라) 한국 무대에서 가장 좋은 성과를 냈던 B형은 가장 나쁜 성과를 내는 데 그쳤다.

이 진술이 사실에 부합하려면 ‘삼진은 능력’이라는 사실을 한 번 더 검증할 필요가 있다. 마찬가지로 이들이 AAA에 복귀한 이후 기록한 성적을 토대로 ‘9이닝당 삼진+’를 계산한 뒤 ANOVA를 실시했다. 삼진을 많이 잡을수록 좋은 투수라 이 기록 역시 높을수록 좋다.

<표 11> 투수 유형별 AAA 복귀 후 9이닝당 삼진+

유형	인원	평균	표준편차	비고
A	11	103.2	23.3	F 3.932 p .010
B	26	97.9	75.1	
C	16	88.0	36.8	
D	12	119.8	28.0	

한국에서 삼진 능력이 좋았던 A와 D가 AAA 복귀 후에도 가장 좋은 성적을 거둔 사실을 90% 신뢰 수준($p=.010$)에서 확인할 수 있다. A형 투수는 구단에서 계약을 거부하는 게 아니라 본인이 메이저리그에 대한 꿈을 좇아 미국으로 돌아가는 일도 많다는 점을 감안하면 D형은 더욱 돋보이는 성적을 남겼다.

이제 확실히 이렇게 강조할 수 있다. 행운이 시간과 장소(리그)에 따라 달라지는 것보다 능력이 훨씬 적게 변한다. 행운이 쌓인다고 실력이 되는 건 아니다. 맞혀 잡는 건 운이지만 삼진은 실력이다. 문제는 볼넷 역시 실력이라는 점이다.

5. 외국인 투수 총 시즌 예측 의사결정나무 모델

이번 연구에서 능력에 집중하고 있는 능력이 '미래 성적'을 담보하기 때문이다. 그래야 외국인 투수와 재계약하는 게 의미가 있다. 이번 해 잘했다는 선물이 아니라 다음 해에 (더) 잘해달라는 기대로 외국인 투수와 재계약하는 것이기 때문이다. 여러 번 재계약에 성공한 투수일수록 더 좋은 투수라고 할 수 있는 이유다.

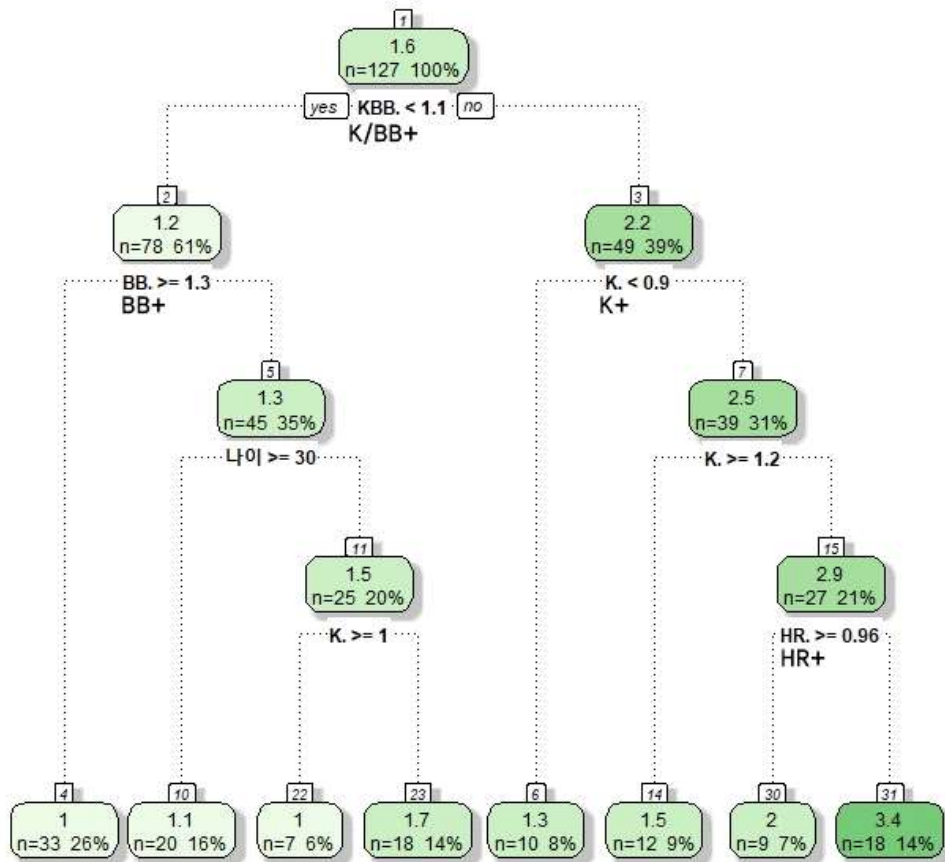
이를 알아볼 수 있도록 이번에는 의사결정나무 분석을 실시하면서 종속변수에 '총시즌'을 넣었다. 이때 총시즌은 이 투수가 한국 프로야구에서 몇 년간 활약했는지를 뜻한다. 첫 계약 팀과 재계약에 실패하고 다른 팀과 재계약한 경우에도 연구 대상에 포함시켰다는 의미다.

첫 계약팀에서는 재계약하지 않은 투수를 다른 팀에서 찾는다는 건 그 팀에서 이 투수가 한국 무대에서 통할 것이라고 예상했다는 뜻이다. 또 어떤 팀에서든 실제로 오래 뛰었다는 건 그 예측이 맞아 떨어졌다는 뜻이라고 할 수 있다.

독립변수에는 키, 몸무게와 함께 삼진, 볼넷, 피홈런을 투입했다. 이번에는 진짜 능력을 알아보는 차원인 만큼 '기록+' 그러니까 '9이닝당 삼진(K)+', '9이닝당 볼넷(BB)+', '9이닝당 피홈런(HR)+'를 활용했다. 외국인 투수들이 한국 무대에 들어온 연도별로 경기당 평균 득점 환경이 다르기에 이를 보정한 것이다. '삼진 대 볼넷 비율(K/BB)'도 'K/BB+' 형태로 넣었다.

그 결과 K/BB+가 1.1 이상일 때, 즉 삼진 대 볼넷 비율이 리그 평균보다 10% 이상 높을 때 한국 프로야구 무대에 머문 총 시즌이 높아지는 올라가는 것으로 나타났다. 결국 K/BB+가 좋을수록 한국 프로야구에서 잘 통하는 투수라고 할 수 있다.

문제는 이 결과를 외국인 선수가 국내 무대에서 첫 시즌을 보내고 나서야 알게 된다는 점이다. 재계약 때는 문제가 없지만 외국인 투수를 새로 영입할 때는 이 기록을 미리 예측할 수 있어야 한다. 다행스러운 건 III-8에서 살펴본 것처럼 삼진과 볼넷 능력은 리그가 바뀌어도 예측 가능한 범위에서 변한다는 점이다.



[그림 16] 외국인 투수 총 시즌 예측 의사결정나무 모델⁴³⁾

6. 삼진 대 볼넷 변화 비교

그래도 삼진 대 볼넷 비율(K/BB)은 변수 두 가지가 동시에 변한다는 점은 감안해야 한다. 삼진과 볼넷이 어떤 방향으로 변하는지에 따라 K/BB 값이 달라질 수 있다.

이 둘은 같은 방향으로 변할 수도 있지만 꼭 그런 것만은 아니다. 그러니까 어떤 투수는 AAA에서 한국 프로야구 무대로 건너오면 삼진이 늘고 볼넷이 줄지만 다

43) R에서는 변수 이름에 '+', '/' 기호를 사용할 수 없어 '.'으로 나왔다.

른 투수는 볼넷이 늘고 삼진이 줄 수도 있는 것이다.

방향뿐 아니라 변화량도 문제가 될 수 있다. 만약 두 기록이 서로 반대 방향으로 변한다고 했을 때 한 기록이 반대 방향으로 가는 다른 기록 변화를 상쇄할 만큼 영향력이 크다면 이 기록이 좀 더 중요한 기록이라고 할 수 있다. 종합하자면 삼진과 볼넷 변화를 무조건 일대일로 생각할 필요는 없다는 뜻이다.

삼진과 볼넷 중 어느 쪽이 K/BB 변화에 더 큰 영향을 끼치는지 알아보려고 ‘K/BB+ 변화량’을 종속변수로, ‘9이닝당 삼진(K)+ 변화량’과 ‘9이닝당 볼넷(BB)+’ 변화량을 독립변수로 선형 회귀분석을 실시했다. 여기서 변화량은 ‘AAA 기록 - 한국 프로야구 기록’을 뜻한다.

그 결과 설명력이 아주 높은 편은 아니지만($R^2=.345$) K+ 변화량($\beta=.584$)이 BB+ 변화량($\beta=-.444$)보다 K/BB+ 변화량에 끼치는 영향이 더 크다는 사실을 확인할 수 있다.

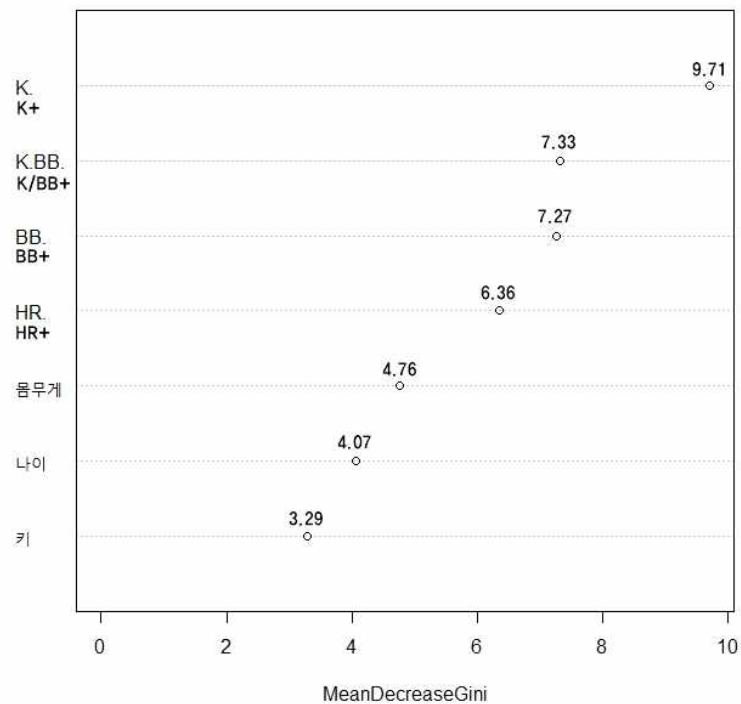
<표 12> K/BB+ 변화량 회귀 분석 결과

종속 변수	독립변수	표준오차	β	t	유의확률
	상수	.055		2.239	.027
K/BB+ 변화량	K+변화량	.202	.584	6.900	.000
	BB+변화량	.130	-.444	-5.239	.000
	R .588, R^2 .345, 수정된 R^2 .333, F=27.939, p=.000				

삼진이 중요하다는 건 AAA 기록을 토대로 이듬해 재계약 여부를 예상하는 ‘랜덤포레스트’ 모델⁴⁴⁾을 통해서도 확인할 수 있다.

종속변수에 재계약 여부를 넣고 독립 변수에 K+, BB+, K/BB+, 나이, 몸무게, 키를 투입하면 K+가 지니 지수 감소량이 가장 큰 변수가 된다. CART 알고리즘에서는 이 숫자가 클수록 중요한 분류 기준점이다.

44) 이 모델은 OOB 오류(Out-of-bag error)가 25.2%에 달해 좋은 모델이라고 보기는 어렵다. 다만 현실에서 외국인 투수를 잘못 판단한 비율이 74.6%(100% - 25.4%)라는 점을 고려하면 참고할 수는 있다.



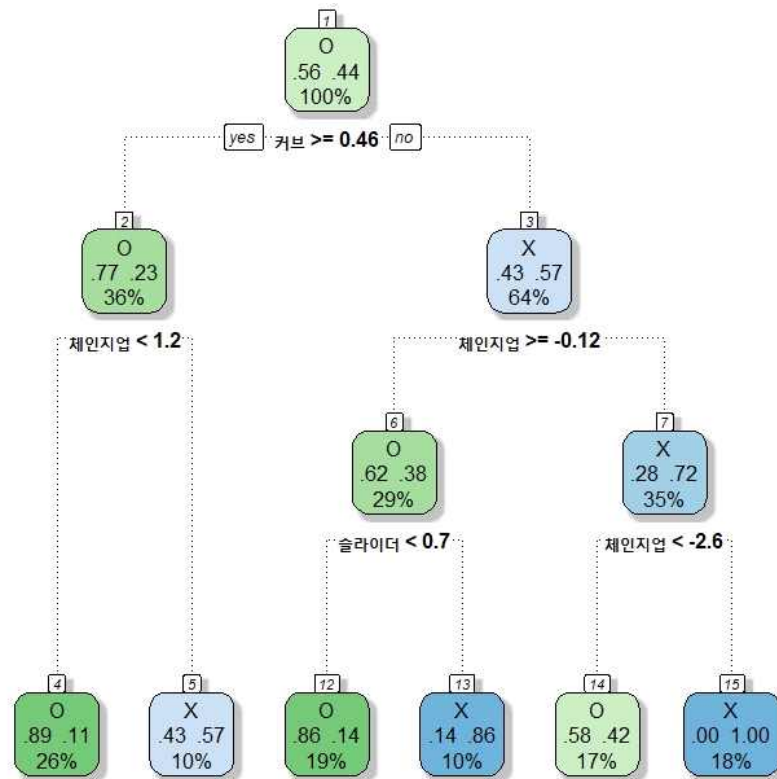
[그림 17] 외국인 투수 재계약 확률 예측 랜덤 포레스트 모델 변수 중요도

우리가 삼진 영향력을 측정한 건 한국 무대에서 높은 K/BB를 기록할 수 있는 투수 특성을 찾으려는 목적이었다. 볼넷 능력은 그대로 둔 상태에서 K/BB를 끌어올리려면 AAA보다 한국 프로야구에서 삼진을 더 많이 잡을 수 있는 투수 유형을 찾으려면 된다.

7. 외국인 투수 구종 가치 의사결정나무 모델

한국 프로야구 진출 직전 AAA에서 활약한 112명 중 한국에 와서 ‘9이닝당 탈삼진(K)+’가 더 좋아진 선수는 53명(47.3%)이다. 이 중 30명(58.8%)이 한국 프로야구에서 구종 가치 기록을 남겼다. 표본 확보 차원에서 중복을 포함하면 이 숫자는 40명으로 늘어난다.

종속변수에 K+ 향상 여부(향상 O, 유지 또는 하락 X)를 넣고 독립변수에 8개 구종별 가치를 넣어 의사결정나무 분석을 실시하면 [그림 15]와 같은 결과를 얻을 수 있다.



[그림 18] 리그 변화에 따른 K+ 향상 예측 의사결정나무 모델

이를 통해 커브가 K+ 향상에 있어 중요한 구실을 담당하고 있는 구종이라는 사실을 확인할 수 있다. 커브 구종 가치가 0.46 이상이면 AAA보다 한국 프로야구에서 K+가 좋아질 확률이 77%로 올라간다.

커브가 중요하다는 건 독립표본 T검정 실시결과로도 확인이 가능하다. K+ 향상여부를 집단변수, 구종별 ‘구종 가치/100’을 입력변수로 놓고 평균을 비교하면 커브만이 99% 신뢰 수준에서 유의미한 차이를 나타낸다.

〈표 13〉 K+ 향상 여부에 따른 커브 구종 가치/100

기록	향상	N	평균	표준편차	t값	p값
K+	O	40	.0526	1.21	-2.87	0.006
	X	32	-.9446	1.71		

뿐만 아니다. 커브는 신뢰도 95% 수준에서 외국인 투수 계약 갱신⁴⁵⁾ 여부에 영향을 끼치는 유일한 구종이기도 하다. 〈표13〉은 집단변수에 재계약 여부, 입력변수에 구종별 ‘구종 가치/100’ 을 넣어 독립표본 T검정을 실시한 결과다.

〈표 14〉 2014~2016년 외국인 투수 커브 구종 가치/100

계약 갱신	N	평균	표준편차	t값	p값
O	19	.0826	2.14	-2.21	.032
X	53	-1.3679	3.18		

결국 커브를 잘 던지는 투수가 한국 프로야구 무대로 건너오면 삼진이 늘어 K/BB+가 올라가고, 이를 통해 계약을 연장할 확률도 끌어올릴 수 있다고 결론 내릴 수 있다.

VI. 결론

1. 한국과 미국 커브는 다르다.

커브는 가장 기본으로 손꼽히는 변화구다. 그런데도 커브가 외국인 투수 성공에 이렇게 큰 영향을 끼치는 이유는 뭘까. 기본적으로 ‘미국 커브’와 ‘한국 커브’는 똑같은 공이 아닐지도 모른다는 연구 결과에서 힌트를 찾을 수 있다.

신동윤 한국야구학회 데이터분과장이 2015년 트랙맨 데이터를 분석한 결과에

45) 다른 팀과 계약해 한국 프로야구에 계속 남는 경우 포함

따르면 2015 한국 프로야구 투수들이 던진 커브의 릴리스 각도는 2.1도인데 2014년 AAA 평균은 0.8도다. 신동윤은 “커브의 릴리스 각도가 낮을수록(속구와 비슷할수록) 타자의 헛스윙 %가 증가한다”고 설명하고 있다(일간스포츠, 2016b). 볼이 헛스윙으로 바뀌면 당연히 볼넷도 줄어든다.



[그림 19] 한국 프로야구와 AAA 커브 릴리스 각도 차이(일간스포츠, 2016b)

게다가 곧바로 ‘2스트라이크 이후 헛스윙률’ 역시 외국인 투수가 던진 커브 쪽이 3% 포인트 높았다.⁴⁶⁾ 2스트라이크 이후에 헛스윙이 나오면 곧바로 삼진이다. 즉 외국인 투수에게 커브는 삼진을 잡는 공이기도 하다. 볼넷이 줄고 삼진이 많아지면 K/BB가 올라간다. 또 국내 선수가 던진 커브를 때렸을 때 장타력은 평균 0.504지만 외국인 투수가 던진 커브를 쳤을 때는 0.389로 내려간다.

재미있는 건 한국 무대로 오는 외국인 투수가 유독 커브 실력이 빼어난 이들은 아니라는 점이다. 신동윤은 “그들(한국에 오는 투수들)이 던지는 커브는 뛰어난 투수만 던질 수 있는 특별한 공으로 보기 어렵다. 구질(각도)과 효과성(헛스윙률)에서 AAA 평균에 해당하는 보통 커브에 더 가깝다. 즉, ‘좋은 커브’와 ‘나쁜 커브’의 차이가 아니라 ‘미국 커브’와 ‘한국 커브’의 차이일 수 있다는 것이

46) 신동윤은 구체적인 숫자는 제시하지 않았다.

다” 고 덧붙였다.

이를 통해 구종 가치가 성과 지표라고 해서 커브의 가치를 허투루 보면 곤란하다는 사실을 알 수 있다. 이런 성과를 내는 밑바탕에 ‘한국에서 보기 힘든 커브를 던질 줄 안다’ 는 능력이 깔려 있기 때문이다. AAA에서는 흔한 커브가 한국 프로야구 무대에서 장수(長壽)를 보장하는 ‘비밀 무기’ 가 될 수도 있는 것이다. 요컨대 ‘미국 커브’ 는 아직 한국 야구에서 저평가된 자원이라고 말할 수 있다.

2. 키 큰 투수는 BABIP를 억제 할까

이유를 찾기 힘든 일도 있었다. 키가 189cm 이상인 외국인 투수는 평균 1.8년 동안 한국 프로야구 무대에서 뛰었다. 5명 중 4명꼴로 재계약에 성공한 셈이니까 전체 평균보다 재계약이 3.2배 높은 것이다.⁴⁷⁾

그 이유를 알아보려고 독립표본 T검정, 회귀분석, ANOVA 등을 진행했지만 키는 삼진(K+)이나 볼넷 억제 능력(BB+), K/BB+ 등과 유의미한 상관관계를 보이지 않았다. 다만 키 189cm 이상일 때 BABIP+가 90% 신뢰 수준에서 유의미하게 낮다는 사실을 확인할 수 있었을 따름이다.

<표 15> 키 189cm 기준 BABIP 차이

기록	키 189cm 이상	N	평균	표준편차	t값	p값
BABIP+	X	76	102.3	26.6	1.826	0.070
	O	61	95.6	12.9		

BABIP는 기본적으로 행운이 개입하는 성과 지표이기 이 같은 결과는 ‘우연’의 산물일 가능성이 높다. Gassko(2006) 역시 제2차 세계대전 이후 메이저리그에서 활약한 투수를 대상으로 조사한 결과 키와 BABIP 사이에 상관관계가 있다고 보기

47) 2018년 1월 12일 현재 각 구단이 계약을 마친 새 외국인 투수 8명 평균 신장이 190.5cm라는 사실 역시 각 구단이 키 큰 외국인 투수를 선호한다는 사실을 증명한다.

는 어렵다고 지적했다.

단, 키가 구종에 끼치는 영향 등에 관해서는 추가 연구가 필요하다고 할 수 있다.⁴⁸⁾ Greenberg(2007)는 키와 유망주가 메이저리거로 성장할 확률 사이에 긍정적인 상관관계가 있다는 사실을 밝혀냈다. 키와 투구 실력 사이에 어떤 식으로든 연관이 있을 수 있다는 사실을 암시하는 대목이다.

PFX 같은 투구추척시스템이 등장하기 전에는 이에 관해 실증적인 연구를 하기가 쉽지 않았지만 이제는 이런 연구를 할 수 있는 가능성이 열린 상태다. 한국 프로야구와 AAA 모두 아직은 측정 기간이 길이 않아 이번 연구에서는 충분한 표본을 확보하지는 못했다.

그럼에도 트랙맨 등 구종 가치를 활용하면 아직 밝혀지지 않았던 야구 지식에 다가갈 수 있다는 사실을 일정 부분 증명한 것 역시 이번 연구의 시사점 가운데 하나라고 할 수 있다.

3. 향후 연구 제안

세이버메트릭션이 아닌 이들에게도 세이버메트릭스 붐을 일으킨 책 ‘머니볼’의 부제는 ‘불공정한 게임을 승리로 이끄는 과학(Moneyball: The Art of Winning an Unfair Game)’이다.

이때 불공정하는 건 이 책이 분석한 메이저리그 구단 오클랜드가 보스턴 같은 ‘빅 마켓 팀’과 비교할 때 선수 영입에 많은 돈을 쓰기가 어렵다는 점이었다. 그러면 당연히 좋은 전력을 갖추기가 어려웠다.

대신 오클랜드가 주목했던 건 당시까지 저평가받고 있던 ‘출루율’이라는 기록이었다. 당시까지는 메이저리그 선수 시장에서 출루율이라는 가치가 높게 평가받지 않던 분위기였기 때문에 오클랜드는 저렴한 선수 몸값을 지불하면서도 꾸준히

48) 일단 이번 연구에서는 구종 가치와 관련한 변수 역시 키와 통계적으로 유의미한 상관관계를 나타내지 않았다.

좋은 성적을 낼 수 있었다.

물론 이런 발견이 영원히 가치가 있을 것이라는 보장은 없다. 오클랜드는 빅마켓 팀도 출루율에 주목하기 시작하면서 상대 우위를 잃었다. 저평가주를 먼저 발견한 건 사실이지만 결국 출루율을 둘러싸고도 ‘불공정한 게임’이 벌어지자 한 발 뒤로 물러설 수밖에 없던 것이다. 세이버메트릭스를 통해 처음 메이저리그 정상을 차지한 건 오클랜드가 아니라 세이버메트릭스와 ‘돈’을 동시에 갖춘 보스턴이었다.

한국 프로야구에서 커브볼을 잘 던지는 선수가 재계약할 확률이 높다는 것 역시 구단에서 투구 데이터에 주목하고 있다는 사실을 방증한다. 단지 구종뿐만이 아니다. 삼성과 SK는 2018년을 함께 할 외국인 투수 계약 소식을 알리면서 나란히 보도자료에 ‘익스텐션(extension)’⁴⁹⁾이라는 ‘트랙맨’ 용어를 사용했다.

“특히 익스텐션 측면에서 공을 최대한 앞으로 끌고 나가 던지는 스타일이기 때문에 공의 무브먼트가 뛰어난 편이다.”⁵⁰⁾

“또한 투구 시 하체 활용과 익스텐션이 우수하여 타자들의 체감 구속이 실제 구속보다 더 빠른 스타일”⁵¹⁾

한국 구단 역시 세이버메트릭스를 통해 외국인 투수 영입 성공 가능성을 높이려 애를 쓰고 있다는 증거다. 행운이 찾아오지 않을 때는 직접 돈보기를 들고 찾아나서는 것도 분명 행운을 찾는 한 방법이다.

다시 한 번 강조하자면 이런 시도가 당장 성공을 이끈다고 안심해서는 안 된다. 나중에는 모두가 같은 조건에서 경쟁을 벌이게 되기 때문이다. 만약 각 팀 타자들이 커브볼을 잘 때려내는 스윙 매커니즘을 집중적으로 연습한다면 커브볼의 가치

49) 투수가 공을 던질 때 투수판으로부터 몸을 끌고 나온 거리

50) 팀 아델만 영입 보도자료. 삼성 홈페이지 bit.ly/2jf1G9b

51) 양헬 산체스 영입 보도자료 SK 홈페이지 bit.ly/2jd7LTE

가 현재와 같을 수가 없다.

따라서 중요한 건 항상 변화에 주목하고 어떤 가치가 숨어 있는지 발견하려는 노력이라고 생각한다. 이런 노력이 계속될 때 평균자책점처럼 눈에 보이는 결과만 좇다가 진짜 훌륭한 자원을 놓치는 실수로부터 프로야구 구단을 보호할 수 있을 것이라고 믿는다.

그런 점에서 스탯캐스트는 외국인 투수뿐 아니라 야구 분석 전반에 걸쳐 새로운 가능성을 열어줄 것이다. 스탯캐스트 데이터를 활용한 새로운 발견이 뒤따르기를 기대한다.

VII. 참고문헌

- 서민구(2014), R을 이용한 데이터 처리&분석 실무, 길벗
- 이상희(2017), KBO 리그 외국인투수들의 마이너 리그(AAA) 성적과 국내 성적 비교, 명지대 기록정보과학전문대학원
- 일간스포츠(2016a), 외국인 선수 평균 연봉, KBO가 NPB 보다 높다, 2016년 1월 21일
- 일간스포츠(2016b), [베이스볼인플레이] '한국 커브'와 '미국 커브'는 다르다, 2016년 4월 6일
- 일간스포츠(2017년 5월 17일), [프로야구단 경영 분석]삼성, 지난해 선수단 비용 100억원 감소
- 유진은(2015), 랜덤 포레스트. 교육평가연구, 28, 427-448.
- 정호운 · 이상우(2015), 프로야구, 가치를 재발견하다, 유진투자증권 연구보고서
- 조선일보(2017 5월 11일), NC가 뽑는 외국인, 왜 항상 대박일까
- 최종후, 서두성(1999), 데이터마이닝 의사결정나무의 응용. 통계청 「통계분석연구」 제4권 제1호(99.봄) 61-83
- Baumer, B., Zimbalist, A.(2013). The sabermetric revolution: Assessing the growth of analytics in baseball. University of Pennsylvania Press.
- Berry, M. J. A. and Linoff, G. S. (1997). Data Mining Techniques, New York : John Wiley & Sons, Inc..
- Bradbury, J. C. (2015). Another Look at DIPS, The Hardball Times
- Breiman L.,(2001). Random Forests, Machine Learning, Vol. 45, 2001, pp. 5-32.
- Gassko, David(2006). Does Size Matter (Part 5), The Hardball Times
- Greenberg, Glenn P.(2007), Does a Pitcher's Height Matter?, Fall 2010 Baseball Research Journal
- Lindbergh, Ben(2015), Mind the Gap: Has Baseball's Minors-to-Majors Jump Really Gotten Harder for Hitters?, Grantland, grantland.com/the-triangle/mlb-preview-exploring-minor-league-hitters-major-league-struggles
- McCracken, Voros(2001), Pitching and Defense, Baseball Prospectus, www.baseballprospectus.com/article.php?articleid=878
- Nathan, Alan M(2012), Determining Pitch Movement from PITCHf/x Data, webusers.npl.illinois.edu/~a-nathan/pob/Movement.pdf
- Quinlan, J. R. (1993), C4.5 Programs for machine learning, San Mateo : Morgan Kaufmann.

- Schoenfeld, David (2012). What we talk about when we talk about WAR, ESPN
- Silver, N.(2012). The signal and the noise: why so many predictions fail--but some don't. Penguin.
- Tango, T. M., Lichtman, M. G., & Dolphin, A. E.(2007). The book: Playing the percentages in baseball. Potomac Books, Inc..
- Varian, H. R.(2014). Big data: New tricks for econometrics. The Journal of Economic Perspectives, 28(2), 3-27.
- Verikas, A., Gelzinis, A., Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. Pattern Recognition, 44(2), 330-349.
- Weinberg, Neil(2017), Sample Size, Fangraphs, www.fangraphs.com/library/principles/sample-size/
- クジラ飛行機(2016). Pythonによるスクレイピング&機械学習 開発テクニック BeautifulSoup,scikit-learn,TensorFlowを使ってみよう, ソシム

IX. 부록

〈표 16〉 연구 대상 명단

이름	연도	구단	출생연도	신체조건	재계약	총 시즌
고든	2011	SK	1978	183cm · 86kg	×	2
곤잘레스	2010	LG	1983	188cm · 98kg	×	1
구토슨	2009	KIA	1977	188cm · 83kg	×	1
글로버	2009	SK	1976	196cm · 90kg	○	3
나이트	2009	삼성	1975	183cm · 79kg	○	6
니코스키	2009	SK · 두산 ⁵²⁾	1973	188cm · 83kg	×	2
니퍼트	2011	두산	1981	201cm · 97kg	○	7
다카쓰	2008	히어로즈	1968	183cm · 81kg	×	1
더마트레	2010	LG	1981	188cm · 95kg	×	1
데이비스	2008	KIA	1975	190cm · 87kg	×	1
데폴라	2010	한화	1982	183cm · 81kg	○	2
디아즈	2008	KIA	1980	185cm · 81kg	×	1
라라	2016	SK	1988	185cm · 81kg	×	1
라미레즈	2012	KIA	1979	185cm · 99kg	×	1
라이트	2010	KIA	1982	193cm · 122kg	×	1
레딩	2010	삼성	1978	180cm · 104kg	×	1
레온	2016	삼성	1988	183cm · 97kg	×	1
레이	2008	SK	1974	188cm · 90kg	×	1
레이번	2007	SK	1974	193cm · 95kg	○	2
레이어	2008	두산	1977	188cm · 92kg	×	1
레이예스	2013	SK	1984	188cm · 104kg	○	2
레일리	2015	롯데	1988	190cm · 90kg	○	3
로드리게스	2007	KIA	1972	185cm · 77kg	×	1
로드리게스	2013	삼성	1987	193cm · 113kg	×	1
로마노	2007	SK	1972	188cm · 88kg	×	1
로위	2016	kt	1984	180cm · 81kg	×	1
로저스	2015	한화	1985	190cm · 90kg	○	2
로페즈	2009	KIA	1975	190cm · 83kg	○	4

52) SK에서 방출한 뒤 같은 해 두산과 재계약

이름	연도	구단	출생연도	신체조건	재계약	총 시즌
루카스	2015	LG	1985	188cm · 92kg	×	1
리마	2008	KIA	1972	188cm · 77kg	×	1
리오단	2014	LG	1986	193cm · 90kg	×	1
리즈	2011	LG	1983	188cm · 90kg	○	3
린드블럼	2015	롯데	1987	193cm · 108kg	○	3
마리몬	2016	kt	1988	185cm · 88kg	×	1
마리오	2012	SK	1984	188cm · 95kg	×	1
마야	2014	두산	1981	183cm · 97kg	○	2
마에스트리	2016	한화	1985	180cm · 79kg	×	1
마틴	2014	삼성	1983	193cm · 99kg	×	1
매그레인	2011	SK	1978	188cm · 90kg	×	1
매존	2007	삼성	1976	188cm · 99kg	×	1
매클레리	2008	롯데	1974	196cm · 99kg	×	1
매티스	2011	삼성	1983	190cm · 104kg	×	1
맥그레거	2016	넥센	1986	193cm · 90kg	×	1
바우어	2009	LG	1977	198cm · 97kg	×	1
바티스타	2011	한화	1982	196cm · 86kg	○	3
배스	2012	한화	1982	188cm · 97kg	×	1
밴덴헐크	2013	삼성	1985	196cm · 97kg	○	2
밴와트	2014	SK	1986	190cm · 99kg	○	3
밴헤켄	2012	넥센	1979	190cm · 83kg	○	6
번사이드	2010	넥센	1977	190cm · 95kg	×	1
벨레스터	2016	삼성	1986	193cm · 86kg	×	1
보우덴	2016	두산	1986	190cm · 97kg	○	2
볼스테드	2014	두산	1986	203cm · 106kg	×	1
부시	2012	SK	1979	188cm · 92kg	×	1
부예노	2010	한화	1981	180cm · 92kg	×	1
부척	2011	롯데	1978	196cm · 95kg	×	1
빌로우	2013	KIA	1985	190cm · 95kg	×	1
사도스키	2010	롯데	1982	193cm · 83kg	○	3
서캠프	2016	한화	1987	196cm · 99kg	×	1
세데뇨	2009	두산	1983	185cm · 90kg	×	1

이름	연도	구단	출생연도	신체조건	재계약	총 시즌
세드리	2007	한화	1978	188cm · 99kg	×	1
세든	2013	SK	1983	193cm · 99kg	○	3
션	2008	삼성	1977	193cm · 104kg	×	1
션헨	2012	한화	1981	190cm · 106kg	×	1
소사	2012	KIA	1985	185cm · 95kg	○	6
스와잭	2015	두산	1985	193cm · 97kg	×	1
스코비	2007	KIA	1979	185cm · 88kg	×	2
스튜어트	2015	NC	1986	188cm · 92kg	○	2
스틴슨	2015	KIA	1988	193cm · 95kg	×	1
시스코	2015	kt	1983	208cm · 122kg	×	1
아담	2013	NC	1987	188cm · 81kg	×	1
애킨스	2009	롯데	1977	183cm · 90kg	×	1
앤서니	2012	KIA	1982	193cm · 102kg	○	2
앨버스	2014	한화	1985	185cm · 90kg	×	1
얀	2008	SK	1975	193cm · 115kg	×	1
어센시오	2014	KIA	1983	188cm · 81kg	×	1
어윈	2015	kt	1987	190cm · 95kg	×	1
에니스	2008	삼성	1979	196cm · 99kg	×	1
에르난데스	2009	삼성	1978	190cm · 83kg	×	1
에반	2015	KIA	1983	183cm · 97kg	×	1
에서튼	2007	KIA	1976	185cm · 88kg	×	1
연지	2009	한화	1977	196cm · 99kg	×	1
오넬리	2011	한화	1983	188cm · 90kg	×	1
오버몰러	2008	삼성	1976	188cm · 88kg	×	1
오카모토	2010	LG	1974	183cm · 94kg	×	1
옥스프링	2007	LG	1977	185cm · 88kg	○	5
울슨	2013	두산	1983	185cm · 92kg	×	1
울프	2014	SK	1982	183cm · 86kg	×	1
웨버	2014	NC	1984	188cm · 92kg	×	1
웹스터	2016	삼성	1990	188cm · 86kg	×	1
윌슨	2007	삼성	1976	193cm · 99kg	×	1
유먼	2012	롯데	1979	193cm · 99kg	○	4

이름	연도	구단	출생연도	신체조건	재계약	총 시즌
이브랜드	2013	한화	1983	185cm · 106kg	×	1
저마노	2011	삼성	1982	188cm · 95kg	×	2
존슨	2009	LG	1982	190cm · 77kg	×	1
주키치	2011	LG	1982	196cm · 92kg	○	3
지크	2016	KIA	1989	196cm · 90kg	×	1
카도쿠라	2009	SK	1973	190cm · 90kg	○	3
카리대	2013	삼성	1983	178cm · 88kg	×	1
카스티요	2016	한화	1989	185cm · 106kg	×	1
카페안	2010	한화	1981	193cm · 111kg	×	1
켈리	2015	SK	1988	188cm · 86kg	○	3
코르테스	2008	롯데	1973	180cm · 88kg	×	1
코리	2011	롯데	1973	178cm · 77kg	×	1
코엘로	2016	넥센	1984	196cm · 113kg	×	1
코프랜드	2016	LG	1987	190cm · 99kg	×	1
콜론	2010	KIA	1979	196cm · 111kg	×	1
쿠비얀	2008	SK	1972	188cm · 74kg	×	1
크루세타	2009	삼성	1981	188cm · 97kg	○	2
클레이	2014	한화	1988	188cm · 81kg	×	1
클로이드	2015	삼성	1987	190cm · 95kg	×	1
타투스코	2014	한화	1985	196cm · 90kg	×	1
탈보트	2012	삼성	1983	185cm · 81kg	×	2
토마스	2008	한화	1977	193cm · 106kg	○	2
토마스	2014	KIA	1984	190cm · 99kg	×	1
트레비스	2011	KIA	1982	190cm · 92kg	×	1
티포드	2014	LG	1984	183cm · 72kg	×	1
페르난도	2011	두산	1982	183cm · 99kg	×	1
프록터	2012	두산	1977	185cm · 88kg	×	1
플란데	2016	삼성	1986	188cm · 81kg	×	1
피가로	2015	삼성	1984	188cm · 80kg	×	1
피노	2016	kt	1983	183cm · 86kg	×	1
피어밴드	2015	넥센	1985	190cm · 102kg	○	3
해커	2013	NC	1983	185cm · 104kg	○	5

이름	연도	구단	출생연도	신체조건	재계약	총 시즌
헨킨스	2013	두산	1983	193cm · 88kg	×	1
허프	2016	LG	1984	185cm · 95kg	○	2
험버	2015	KIA	1982	190cm · 97kg	×	1
헥터	2016	KIA	1988	188cm · 88kg	○	2
홀튼	2014	KIA	1979	193cm · 106kg	×	1
히메네스	2010	두산	1980	188cm · 88kg	×	1