한국빅데이터학회지
제7권 제1호, 2022, pp. 49-61 https://doi.org/10.36498/kbigdt.2022.7.1.49

한국프로야구에서 장타율과 출루율(OPS) 예측 연구

Prediction of OPS(On-base Plus Slugging) in KBO League

신동윤 · 김진호†

강원대학교 대학원 컴퓨터과학과

요 약

스포츠 분야에서는 팀 전략 구상과 마케팅 등 팀 운영에 있어서, 데이터 분석의 비중이 점점 더 커지고 있다. 특히, 한국프로야구에서는 한 시즌이 끝나면 FA, 트레이드 등 다음 해 팀 전략을 구상하기 위해서 선수 영입과 선수 육성 등의 계획을 수립하는데, 이 때 선수들의 다음 해 성적을 예측하는 것이 매우 중요하다. 본 연구에서는 타자만으로 대상을 한정지어 다음 해의 성적이 상승할지를 예측해보고자 하였다. 상승 및 하락의 기준이 되는 기록으로는, 계산하기 쉽고 팀 득점과의 관계가 높은 OPS로 하였다. 본 연구에서 데이터는 한국프로야구 1982년부터 2021년까지 40년간의 정규시즌 데이터를 사용하였고, 실험 방법으로는 11개의 머신러닝 분류 모델을 사용하였다. OPS의 상승 및 하락 여부를 예측해본 결과, RBF SVM, Neural Net, Gaussian Process, AdaBoost가 다른 분류 모델에 비해 정확도가 높게 나왔고 나이는 정확도에 큰 영향을 주지 못했다.

■ 중심어 : 야구, OPS, 출루율, 장타율, A.I., 인공지능, 머신러닝, 분류

Abstract

In sports, the proportion of data analysis in team management such as team strategy planning and marketing is increasing. In KBO(Korea Baseball Organization) league, in particular, plans such as recruiting players and fostering players are established to devise team strategies for the next year, such as FA and trade, at the end of a season. For these reasons, it is very important to predict players' performance for the next year. In this study, the target was limited to only the batter and tried to find out how to predict whether the performance of the next year will improve. As a standard record for rising and falling, OPS(On-Base Plus Slugging), which is easy to calculate and has a high relationship with team score, was used. In this study, 40 years of regular season data from 1982 to 2021 were used as data, and 11 machine learning classification models were used as experimental methods. Predicting the rise and fall of OPS, RBF SVM, Neural Net, Gaussian Process, and AdaBoost were more accurate than other classification models, and age did not significantly affect accuracy.

■ Keyword : Baseball, OPS, On-base, Slugging, A.I., Artificial Intelligence, Machine Learning, Classification

2022년 05월 06일 접수; 2022년 05월 31일 수정본 접수; 2022년 06월 07일 게재 확정.
* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1F1A1059255).
† 교신저자 (jhkim@kangwon.ac.kr)

Ⅰ. INTRODUCTION

1.1 연구동기

야구에서 득점이란 3아웃이 되어 이닝이 끝나기 전에 주자가 정규로 1루, 2루, 3루, 본루에 닿을 때마다 1점이 기록되는 것을 말하며, 정식경기가 끝났을 때 이 규칙에 따라 더 많이 득점한 팀이 승자가 된다[1].

더 많이 득점을 하기 위해서는 우선 타자는 더 많이 출루를 해야 하고, 주자가 있다면 진루를 시켜야 한다. 그렇기 때문에, 야구에서는 득점에 기여도가 높은 타자를 선별하기 위해 많은 지표들이 사용되어왔고, 또 새로운 평가 모델에 대한 연구도 계속되어왔다.

Blakeley B.McShane 등은 계층적 베이지안 모델을 이용하여 선수별 OBP(On Base Percentage), SLG(SLuGging percentage), OPS(On-base Plus Slugging Percentage) 등의 지표로 선수들의 공격력을 측정하기 위한 방법을 제안하였다[2]. 조영석 등은 상관분석, 군집분석, 회귀분석을 이용하여 출루율과 장타율이 득점에 미치는 영향을 연구하였다[3]. 김혁주는 OBP, SLG, OPS와 득점간의 상관관계에 따라 가중치를 부여하는 OPS를 제안하였다[4]. 정진상은 빅 데이터 기술을 이용하여 “하이볼포인트”라는 타자 평가 모델을 제안하였는데[5], 이 모델은 타자가 투수에게 얼마나 더 많은 공을 던지게 했는지가 평가항목에 포함된다. 정예린은 동일한 타격결과라도 상대 투수의 기량에 따라 다르게 평가할 수 있는 타자 평가 모델을 제안하였다[6]. 김예형은 타율과 출루율에서 희생플라이를 처리하는 방식이 다르다는 점에 주목하여, 현재 사용되고 있는 출루율을 정의하는 식의 분모에서 희생플라이수를 제거하여 수정출루율을 정의하였고, 이를 바탕으로 수정OPS를 정의하였다[7].

타자의 평가모델에 관한 연구는 이와 같이 계속 발전되어 왔다. 이런 평가모델들은 선수들의 고과산정 등에 이용되어 왔는데, 이제는 과거에 대한 평가뿐만 아니라 선수들의 성적 예측에 관한 연구도 활발히 진행되고 있다.

김민택 등은 선수의 성적 예측은, 구단 관계자들에게는 선수들의 연봉 협상 및 팀에 부족한 부분을 채워줄 새로운 선수를 영입하는데 사용할 수 있고, 해설자들은 성적 예측을 통해 시청자들의 경기에 대한 이해도와 흥미를 높이고 경기 몰입에 도움을 주는 등 다양한 방법으로 이용될 수 있다고 했고[12], 홍종선 등은 과거 3개년도 자료를 바탕으로 타자력 예측 모형을 개발하였다[15].

본 연구에서는 타자의 성과를 나타내는 지표 중에서 OPS의 상승 및 하락을 예측하였는데, 조영석 등은 타자의 성과를 측정할 수 있는 지표는 여러가지가 있지만, 특히 OPS는 득점에 가장 높은 상관계를 갖는 기록이라고 했고[3], 박지훈은 OPS는 출루율이 장타율보다 더 득점 생산력이 크다는 점을 반영하지 못하고 값이 직관적이지 못하다는 단점이 있음에도 불구하고 계산이 너무나 간단하면서도 타자의 공격력을 설명하기에 충분한 지표[13]라고 했다.

한국프로야구에서도 OPS 예측에 관한 연구가 활발히 이루어 지고 있는데, 한정섭 등은 XGBoost (Extreme Gradient Boosting) 예측기법은 최고의 OPS 예측 성능을 보여주었다고 했고[14], 김민택 등은 빅데이터 분석 기법을 이용하여 타자의 OPS를 예측해 보았다[12].

하지만, 지금까지 연구된 것은 주로 회귀분석에 의한 OPS 값의 예측에 관한 연구이고, 본 연구는 OPS의 값을 직접 예측하는 것이 아니라 OPS의 상승 및 하락 여부를 예측한 연구이다. 수치까지 정확히 예측할 수 있다면 최고의 예측 방법이겠지만, 수치를 정확하게 예측하는 데에는 한계가 있다. 따라서, 본 연구에서는 예측의 정확도를 높일 수 있다면 OPS 값이 아닌 OPS의 상승 및 하락 여부를 예측하여도 팀의 전략 구상에 도움이 될 수 있을 것이라고 생각하였다. 이러한 목

적으로, 본 연구에서는 머신러닝의 분류 모델을 사용하여 타자들의 OPS 상승 및 하락 여부를 예측해 보고, 어떻게 하면 더 정확하게 예측할 수 있는지를 알아보고자 하였다.

1.2 연구 한계

- 본 연구에서 사용한 데이터는 1982년부터 2021년까지의 한국 프로야구 기록 중에서 정규리그 결과만으로 한정하였다.
- 본 연구는 각 타자의 기록만을 대상으로 하였으며, 해당 연도의 투수 능력이나 소속 팀의 홈 구장 크기, 스트라이크 존, 공의 반발계수 등 기록에 영향을 미칠 수 있는 다른 외부적인 요인들은 고려하지 않았다.
- 본 연구는 수비 위치에 따른 타자의 체력이나 심리적 부담을 고려하지 않았다.
- 본 연구는 선수들의 부상 또는 부상 후유증 등은 고려하지 않았다.

Ⅱ. 이론적 배경

2.1 야구와 데이터

야구는 기록에서 시작해 기록으로 끝나는 스포츠라고 한다. 보통 스포츠 기록은 순위나 승패 등 결과를 알려주는 데 그치지만, 야구 기록은 “과정”까지 보여주기 때문일 것이다. 기록원이 작성한 야구 기록지를 보면, 그 경기를 직접 보지 않더라도 투수가 던진 공과 타자의 타격 결과를 비롯한 1회부터 9회까지의 경기 전체를 재구성할 수 있다[9]. TV 중계가 없던 시절에는 신문으로 이런 기록들을 가지고 경기내용을 전달할 수 있었다.

하지만, 미국 메이저리그에서도 150여 년의 야구 역사 동안 쌓인 엄청난 기록은 경기 결과를 전하고 선수들의 성적을 정리하는 데 주로 이용되었다. 그런데, 1990년대 이후 메이저리그 선수 연봉이 폭등하면서 데이터를 통해 보다 정교하게 선수 가치를 평가 할 필요성이 커졌다. 이에 따라 세이버메트릭스 (야구에 게임 이론과 통계학적 방법론을 적극적으로 접목해 야구의 본질에 대한 학문적인 접근을 시도하는 방법론) 바람이 불게 되었고, 영화 “머니볼”로도 알려진 오클랜드 애슬레틱스의 성공 이후 이제는 야구에서 기록은 단순한 결과가 아니라, 효율성 있게 구단을 운영하는 데 꼭 필요한 기초자료가 된 것이다[9]. 오늘날 야구에서 데이터는 전통적인 훈련, 스카우트, 전력분석처럼 야구의 일부로 자연스럽게 자리 잡았다[9].

이렇게 이제 야구에서 대량의 데이터를 처리하고, 통계 모델을 만들고, 이를 기반으로 미래를 위해 선수를 스카우트하고, 현장과 프런트의 의사결정에 활용하는 것은 모든 프로구단의 주요한 업무가 되었다[10]. 또한 필드에서 라인업을 짜고 수비 시프트를 펼치고, 한 번도 맞대결한 적이 없는 투수와 타자의 상대 결과를 예측하고 대응책을 찾는 것도 가능하다[10].

2.2 OPS의 의미

OPS는 ‘On Base Percentage Plus(+) Slugging Percentage’의 약자로, ‘출루율+장타율’이다. 1984년 존 쏜과 피트 파머에 의해 처음으로 소개되었다. 이후 뉴욕 타임스 및 ESPN 에서 사용되며 서서히 대중들에게 알려지기 시작하였다[18]. 마이클 루이스가 “팀 득점을 설명할 수 있는 가장 좋은 공격지표”라고 까지 극찬한 OPS와 팀 득점 사이의 상관관계는 0.945에 달한다[17].

영화 “머니볼”에서 브래드 피트가 연기한 빌리 빈이 선수들을 영입한 기준도 OPS 데이터인데[16], OPS가 높은 선수는 타점이나 득점 생산력

〈표 1〉 OPS의 정의

출루율	(안타+사사구) / (타수 사사구 희생플라이)
장타율	[단타 (2*2루타) (3*3루타) (4*홈런)] / 타수
OPS	장타율 + 출루율

이 높을 수밖에 없다. 그래서 빌리 빈은 다른 지표보다는 OPS가 높은 선수를 뽑은 것이다.[16]

세이버메트리션들이 고안해놓은 각종 지표들과 비교했을 때 압도적으로 계산하기 편리하다는 장점이 있다. 이후 세이버메트릭스가 발전하면서 구체화된 wOBA나 EqA 등 수많은 지표들과 비교해도 득점 관계율이 크게 밀리지 않는다. 대표적인 예로 OPS+와 wRC+를 비교해보면 거의 차이가 나지 않는다.[18]

2020년부터는 프로야구 롯데의 부산 사직구장 전광판에는 타자 이름 옆에 타율 대신 OPS가 나오는데[19], 이렇게 OPS는 팀과 관중들이 흥미를 갖고 보고 있는 타자 성적 지표이다.

〈표 2〉 연도별 50 타석이상 출전한 타자들의 OPS 분포

	(OPS*1000)/100	선수 수
1	2	5
2	3	58
3	4	248
4	5	809
5	6	1452
6	7	1426
7	8	795
8	9	306
9	10	90
10	11	19
11	12	2

Ⅲ. 데이터 설명

본 연구에서는 스탯티즈[11]에 있는 1982년부터 2021년까지 40년간의 한국 프로야구 정규리그의 데이터를 기초로 하였다. 기록실 > 시즌기록실 > 타격메뉴에 있는 연도별 선수 데이터 9,680건을 MS-SQL 데이터베이스에 저장한 후, 선수의 개명정보를 토대로 데이터를 모두 새 이름으로 Update 하였다. 그렇게 한 후에, 투수가 아니고 50타석 이상 출전한 연도의 데이터만을 갖고, “선수명_생년월일”의 형식으로 ID를 부여하였고, 이렇게 ID를 부여한 선수별-연도별 데이터 중에서 8년이상 출전한 선수들의 데이터 282건으로 실험 데이터를 생성했다.

선수별로 8년간의 데이터가 실험데이터인 것이다. 2021년 기록이 있는 선수라면 2014~2021까지, 2020년까지의 기록만 있다면 2013년~2020년까지의 기록으로 8년간의 데이터를 만들었다. 만약에 2011~2020까지 총10년치의 기록을 가지고 있는 선수가 있는데, 그 중에서 2013, 2018년에 50타석을 채우지 못했다면, 50타석을 채운 8시즌의 데이터로 실험데이터를 만들었다.

스탯티즈[11]의 개명정보 목록에 없는 경우에는 동일한 선수일지라도 각각 다른 선수로 인식하는 한계가 있다.

〈표 3〉 최초 수집한 데이터 형태

이름	생년월일	연도	나이	게임수	OPS
강귀태	1979-08-12	02	23	74	.636
강귀태	1979-08-12	03	24	66	.705
강귀태	1979-08-12	04	25	84	.748
강귀태	1979-08-12	05	26	104	.738
강귀태	1979-08-12	06	27	73	.754
· · · 총 9,680 건					

〈표 4〉 선수별 + 연도별 데이터를 선수별 데이터를 변경한 형태

데이터 확인용		독립데이터														종속 데이터
PLAYER_ID	예측 년도 OPS	-7 년		-6 년		-5 년		-4 년		-3 년		-2 년		-1 년		상승 하락 여부
		AGE	OPS	AGE	OPS	AGE	OPS	AGE	OPS	AGE	OPS	AGE	OPS	AGE	OPS	
강귀태_ 1979-08-12	6	25	7	26	7	27	7	28	5	29	6	30	6	31	5	1
강동우_ 1974-04-20	5	31	6	32	6	34	3	35	8	36	6	37	7	38	6	0
강민호_ 1985-08-18	8	29	7	30	10	31	9	32	8	33	7	34	7	35	8	1
강병식_ 1977-04-23	5	26	5	28	7	29	7	30	7	31	7	32	7	33	8	0
강봉규_ 1978-01-12	5	28	6	29	7	30	7	31	9	32	6	33	6	34	7	0
⋮ 총 282 건																

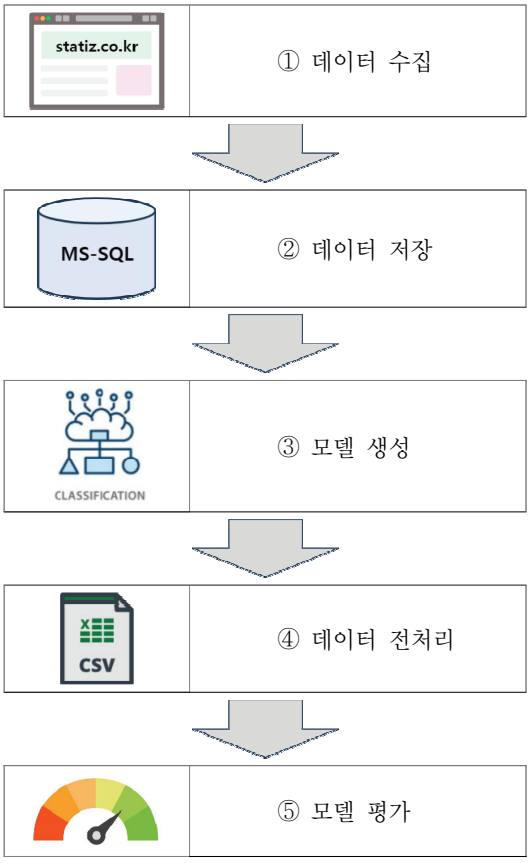
〈표 5〉 실험별로 사용된 데이터 컬럼

		실험에서 컬럼 사용 여부														
실험구분	나이 포함 여부	-7 년		-6 년		-5 년		-4 년		-3 년		-2 년		-1 년		상승 하락 여부
		AGE	OPS	AGE	OPS	AGE	OPS	AGE	OPS	AGE	OPS	AGE	OPS	AGE	OPS	
2년치 데이터로 예측	X												O		O	O
	O											O	O	O	O	O
3년치 데이터로 예측	X										O		O		O	O
	O									O	O	O	O	O	O	O
4년치 데이터로 예측	X								O		O		O		O	O
	O							O	O	O	O	O	O	O	O	O
5년치 데이터로 예측	X						O		O		O		O		O	O
	O					O	O	O	O	O	O	O	O	O	O	O
6년치 데이터로 예측	X				O		O		O		O		O		O	O
	O			O	O	O	O	O	O	O	O	O	O	O	O	O
7년치 데이터로 예측	X		O		O		O		O		O		O		O	O
	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

〈표 6〉 연도별 데이터 건수

연도	타자수	연도	타자수
1982	108	2002	215
1983	117	2003	216
1984	133	2004	224
1985	144	2005	206
1986	172	2006	193
1987	167	2007	277
1988	178	2008	278
1989	190	2009	278
1990	198	2010	269
1991	235	2011	263
1992	235	2012	284
1993	217	2013	257
1994	233	2014	306
1995	227	2015	381
1996	239	2016	378
1997	232	2017	360
1998	202	2018	331
1999	198	2019	356
2000	202	2020	373
2001	214	2021	394
합계			9,680

〈표 7〉 실험 프로세스



Ⅳ. 실 험

4.1 실험방법

다음해의 OPS가 상승할지 하락할지를 머신러닝의 분류 알고리즘을 사용하여 실험해 보았다. OPS 값은 소수점 셋째자리까지 기록하는데, 이진 분류로 OPS를 예측하기 위하여 OPS에 1,000을 곱하고 이 값을 다시 100으로 나누어, OPS가 1자리~2자리 자연수로 표현되도록 하였다. 그리고 나서, 전년도의 값과 비교하여 값이 같거나 상승했으면 1로, 하락했으면 0으로 설정했다. 즉, 2020년에 OPS가 0.751이라면 7로 만들고 2021년의 OPS가 0.691이라면 6으로 만든 뒤에, OPS가 하락했기 때문에 상승 및 하락 여부는 0으로 설정하는 것이다.

그리고, 모델에 데이터를 적용하기 전에 표준화(StandardScaler) 처리를 하여 정규화를 하였다.

실험 결과의 신뢰성을 높이기 위하여 실험은 두 개의 데이터세트를 사용하여 진행하였다. 이 두 개의 데이터세트를 고르기 위해서 모의 테스트를 진행하였다. 모의 테스트는 XGB 모델을 사용하여 2년치 OPS로 다음 해의 OPS를 예측하였는데, XGB 모델은 한정섭[14] 등이 진행한 연구에서 가장 높은 정확도를 보였기 때문에 모의 테스트에서 사용하였다. random state를 최초 10으로 설정하고 10씩 더해가면서 총 10번(10,

20, 30....100) 실행하여 그 10번의 결과 중에서 가장 높은 정확도를 보여주었던 데이터세트(random state:80)를 실험데이터세트1(이하 [SET 1])로, 가장 낮은 정확도를 보여준 데이터세트(random state:70)를 실험데이터세트2(이하 [SET 2])로 구성하였다. 다시 말하면, [SET 1]과 [SET 2]는 소스 데이터는 같지만, 훈련데이터와 테스트데이터로 나눌 때 그 구성을 다르게 한 두 개의 실험데이터세트이다.

실험에서는 우선 2년~7년치의 OPS만을 기준으로 다음 해의 OPS가 유지 또는 상승할지 아니면 하락할지를 예측해 보았다. 다시 말하면, 최초 입력 데이터는 2년치의 OPS이다. 즉, 전년도 OPS와 전전년도 OPS 이렇게 두 개만 입력한 것이다. 그리고 나서, 입력 데이터에 OPS를 1년치씩 추가할수록 정확도가 어떻게 되는지를 살펴보았다. 이렇게 2~7년치의 OPS만을 입력 데이터로 사용해서 실험을 진행한 후, 나이를 입력데이터에 추가하여 다시 한 번 실험을 진행해 보았다.

Mitchel Lichtman[20]은 1950년~2008년의 메이저리그 타자들의 노화곡선(Aging Curve)을 분석한 결과 27~28세가 최고 정점이었다고 했고, 이장택[21]은 연도, 팀의종류, 볼넷, 자유계약선수여부, 나이의 제곱, 안타, WARS, 경험, 타점이 타자의 연봉에 영향을 주는 중요한 변수라고 하였고, 박성배[23] 등은 FA 미대상자 분석 결과, RC/27(한 타자로 1~9번까지 타순을 구성했을 때 한 경기 동안 몇 점을 생산 할 수 있는지를 보여주는 지표), 홈런, 연령 변수가 전체 예측자 중요도의 85%를 차지했다고 하였다. 하지만, 홍종선 등[15]은 대부분의 유의한 타자력 지표들은 연령과 관계가 없다고 하였고, 오상진[22]은 노화곡선(Aging Curve)은, 같은 나이의 선수라도 선수 간의 격차가 큰 것을 고려하지 않고 타자의 평균 성적만을 보여주기 때문에, 이것만으로 선수의 성적에 대해서 단언할 수 없다고 하였다. 이렇게 나이는 모든 선수가 노화 곡선을 갖고 있고 또 연봉 등에서는 중요한 변수이기는 하지만, 아직 타자들의 성적과 나이와 관련된 연구가 미비하기 때문에, 이번 실험에서는 나이와 OPS가 어떤 관계가 있고 이것을 바탕으로 향후 어떻게 나이를 적용할 수 있는지를 알아보기 위하여 나이를 입력데이터에 추가하였다. 나이도 모델에 적용하기 전에 표준화(StandardScaler) 처리를 하였다.

실험에 사용한 분류 모델은 총 11가지로, Nearest Neighbors, Linear SVM, RBF SVM, Gaussian Process, Decision Tree, Random Forest, Neural Net, AdaBoost, Naive Bayes, QDA, XGB 이다.

4.2 실험결과

2년~7년치의 데이터를 입력값으로 하고 다음 해의 OPS의 상승 및 하락 여부를 예측 했을 때, 연도별 OPS 입력 값이 많을 수록 정확도가 높아지는지 확인했다. 실험결과는 [SET 1]과 [SET 2] 모두 2년치의 데이터로 다음해의 OPS 상승 및 하락 여부를 예측하는 것이 3~7년치의 OPS를 입력값으로 예측했을 때보다 가장 정확도(accuracy)가 높게 나온 모델이 많았다. 세부적으로 보면, [SET 1]에서는 2년치의 데이터로 예측했을 때, 나이 미포함의 경우 11개 모델 중 7개의 모델에서, 나이 포함의 경우 11개 모델 중 6개의 모델에서 가장 정확도가 높게 나왔고, 모델별 정확도의 총합도 가장 높았다. [SET 2]의 경우에는 나이 포함 여부에 상관없이 2년치의 데이터로 예측 했을 때 모두 11개의 모델 중 5개의 모델에서 가장 정확도가 높게 나왔다. 11개의 모델 정확도의 총합으로 비교해 보면, 나이가 입력값에 포함되지 않았을 때에는 4년치의 OPS를 입력값으로 했을 때가 가장 높았고, 나이가 입력값에 포함되었을 때에는 2년치의 OPS로 예측하는 것이 가장 높았다.

〈표 8〉 SET 1 (나이 미포함)의 예측 결과

정확도 : accuracy	2년	3년	4년	5년	6년	7년
Nearest Neighbors	0.56	0.75	0.65	0.63	0.61	0.61
Linear SVM	0.42	0.42	0.42	0.42	0.42	0.42
RBF SVM	0.74	0.74	0.72	0.74	0.70	0.67
Gaussian Process	0.74	0.72	0.72	0.75	0.70	0.68
Decision Tree	0.77	0.67	0.63	0.65	0.60	0.61
Random Forest	0.74	0.61	0.70	0.70	0.68	0.70
Neural Net	0.65	0.60	0.72	0.72	0.74	0.70
AdaBoost	0.81	0.68	0.67	0.67	0.63	0.68
Naive Bayes	0.72	0.72	0.72	0.72	0.67	0.70
QDA	0.74	0.72	0.74	0.70	0.70	0.68
XGB	0.81	0.63	0.60	0.58	0.68	0.63

〈표 9〉 SET 1 (나이 포함)의 예측 결과

정확도 : accuracy	2년	3년	4년	5년	6년	7년
Nearest Neighbors	0.68	0.65	0.54	0.53	0.54	0.58
Linear SVM	0.42	0.42	0.42	0.42	0.42	0.42
RBF SVM	0.72	0.70	0.74	0.67	0.65	0.65
Gaussian Process	0.74	0.70	0.70	0.72	0.70	0.70
Decision Tree	0.68	0.68	0.60	0.61	0.65	0.67
Random Forest	0.72	0.65	0.60	0.63	0.65	0.56
Neural Net	0.72	0.70	0.61	0.74	0.60	0.72
AdaBoost	0.68	0.60	0.56	0.61	0.60	0.67
Naive Bayes	0.72	0.63	0.60	0.67	0.67	0.67
QDA	0.63	0.70	0.70	0.63	0.63	0.68
XGB	0.68	0.68	0.68	0.68	0.63	0.70

〈표 10〉 SET 2 (나이 미포함)의 예측 결과

정확도 : accuracy	2년	3년	4년	5년	6년	7년
Nearest Neighbors	0.58	0.61	0.65	0.63	0.70	0.67
Linear SVM	0.49	0.49	0.49	0.49	0.49	0.49
RBF SVM	0.63	0.61	0.63	0.60	0.63	0.61
Gaussian Process	0.58	0.60	0.60	0.61	0.60	0.60
Decision Tree	0.61	0.65	0.65	0.63	0.58	0.61
Random Forest	0.65	0.68	0.67	0.63	0.58	0.61
Neural Net	0.68	0.61	0.63	0.63	0.60	0.60
AdaBoost	0.70	0.67	0.67	0.65	0.68	0.65
Naive Bayes	0.61	0.61	0.61	0.60	0.58	0.56
QDA	0.63	0.58	0.60	0.61	0.63	0.63
XGB	0.61	0.61	0.60	0.68	0.61	0.61

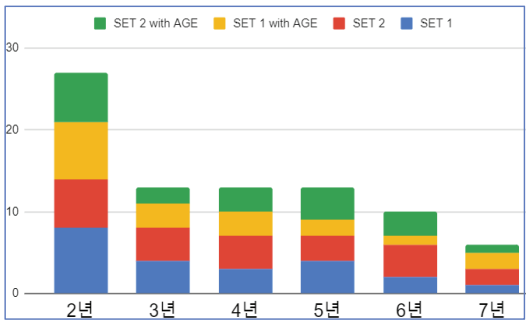
〈표 11〉 SET 2 (나이 포함)의 예측 결과

정확도 : accuracy	2년	3년	4년	5년	6년	7년
Nearest Neighbors	0.58	0.61	0.63	0.67	0.61	0.61
Linear SVM	0.49	0.49	0.49	0.49	0.49	0.49
RBF SVM	0.68	0.67	0.67	0.67	0.68	0.63
Gaussian Process	0.67	0.63	0.60	0.61	0.61	0.61
Decision Tree	0.67	0.63	0.68	0.65	0.63	0.61
Random Forest	0.60	0.58	0.65	0.63	0.72	0.68
Neural Net	0.67	0.60	0.56	0.65	0.65	0.63
AdaBoost	0.70	0.63	0.63	0.65	0.60	0.60
Naive Bayes	0.63	0.65	0.67	0.67	0.65	0.63
QDA	0.60	0.65	0.56	0.51	0.58	0.51
XGB	0.61	0.53	0.54	0.61	0.54	0.47

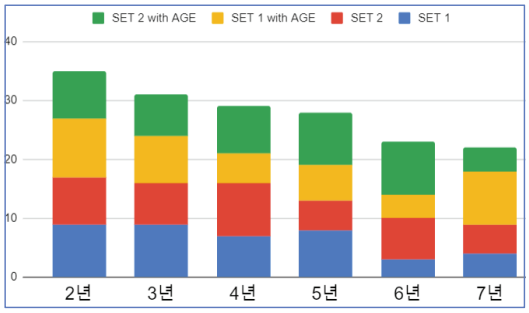
[그림 1]과 같이 2년치의 데이터를 사용했을 때가 [SET 1, 나이 미포함], [SET 2 나이 미포함], [SET 1, 나이 포함], [SET 2 나이 포함] 네 가지 경우를 통틀어서 2년치의 데이터만 입력했을 때가 더 많은 연도의 데이터를 추가로 입력했을 때보다 정확도가 1위였던 경우가 가장 많았다. [그림 2]의 경우에서 보듯이 1~3위까지를 놓고 보면, 그 격차는 많이 줄어들었지만, 그래도 2년치의 데이터만을 사용했을 때가 1~3위 이내의 정확도를 기록하는 경우가 가장 많은 것으로 나왔다.

[SET 1]의 경우에는 2~7년치 모두의 경우에 있어서, 나이를 입력값으로 추가하지 않았을 때가 정확도의 총합이 더 높았다. 하지만, [SET 2]의 경우에는 2, 5, 6년치 데이터를 입력값으로 했을 때에는 나이를 입력값으로 추가하는 것이 정확도의 총합이 더 높았고, 3, 4, 7년치의 경우에는 OPS만을 입력값으로 하고 나이를 입력값으로 하지 않았을 때가 정확도의 총합이 더 높았다.

모델별로 보면, 2년치 데이터만으로 실험을 하였을 때에는 XGB 모델의 정확도가 다른 모델



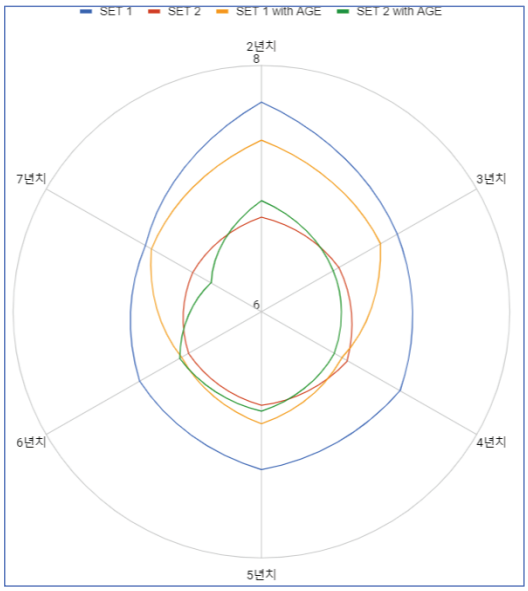
<그림 1> 입력값별 정확도 1위 횟수



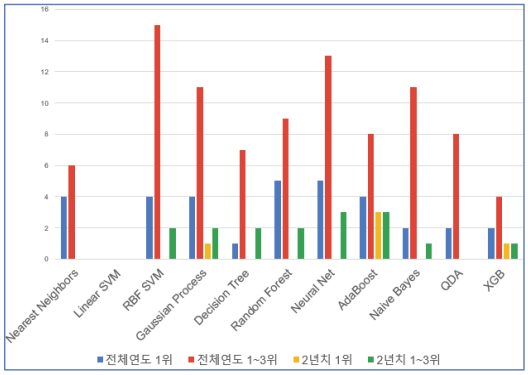
〈그림 2〉 입력값별 정확도 1~3위 횟수

과 비교했을 때 가장 정확도가 높은 모델 중의 하나이었지만, 입력값을 계속 추가하면서 실험했을 때에는 XGB보다 RBF SVM, Neural Net, Gaussian Process, AdaBoost 등의 모델의 정확도가 높게 나왔다.

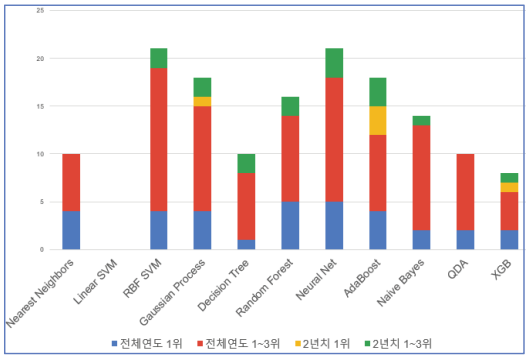
조금 더 자세히 살펴 보면, 2~7년치 모든 경우에 있어서, 정확도 1위를 가장 많이 한 모델은 Random Forest와 Neural Net 이다. 모델별로 정확도 수치가 차이가 크지 않은 경우도 있기 때문에, 얼마나 많이 1~3위에 들었는지로 조건을 변경하면 RBF SVM이 가장 많이 정확도 3위 안에 들었고, Neural Net, Gaussian Process, Naive Bayes 등의 순으로 나타났다. 이것을 OPS 예측이 가장 정확한 2년치 데이터를 입력값으로 했을 때로 한정하면, AdaBoost가 가장 높았고 그 다음으로 Neural Net, RBF SVM, Gaussian Process, Decision Tree, Random Forest 등이 정확도가 높았다.



〈그림 3〉 입력값별 11개 분류모델 정확도의 합



〈그림 4〉 모델별 정확도 1위 / 1~3위 횟수



〈그림 5〉 모델별 정확도 1위 / 1~3위 누적 횟수

그리고, 정밀도, 재현율, F1 Score(정밀도와 재현율의 조화평균)를 확인해 보았다. 2년치 데이터로 예측한 실험에 대해서 살펴보면, 정확도가 가장 높았던 AdaBoost가 F1 Score도 가장 높게 나왔다.

RBF SVM, Gaussian Process, Decision Tree, Random Forest, Neural Net, Naive Baye, QDA, XGB 모두 정확도가 높게 나왔을 때에는 F1 Score도 함께 높았다. 다른 모델에 비해서 정확도가 높았을 경우라도 F1 Score는 낮은 경우가 있는데, F1 Score가 다른 모델에 비해 높았을 경우에는 정확도가 낮게 나온 경우는 없었다. Linear SVM은 재현율이 1로 나오는데, 이것은 모든 실험에서 항상 “1”로 예측했기 때문이다.

Ⅴ. 결 론

요즘 프로야구에서는, 예전에는 감독의 감각에 의해서 운영되었던 것과는 다르게 세이버매트릭스, 빅데이터 분석 등을 활용하여 선수 영입 및 전략 구상을 하고 있다.

다음 해 타자들의 OPS가 상승할지 아니면 하락할지 예측할 수 있다면, FA나 트레이드 등에 의한 선수들의 영입 및 선수 육성 등 팀의 전략 구상에 도움이 될 것이라고 생각하고 이 연구를 시작하였다.

연구 결과, OPS 예측은 2년치의 기록을 입력값으로 했을 때가 가장 정확도가 높게 나타났다. 그리고, 선수별로 노화곡선(Aging Curve)이 존재함에도 불구하고 이번 연구에서는 OPS의 상승 및 하락과 나이와의 관계를 알 수 없었다. 홍종선 등[15]도 대부분의 유의한 타자력 지표들은 연령과 관계가 없다고 하였는데, 이번 연구에서도 나이를 입력값에 추가해도 대부분의 경우 정확도가 향상되지 않았다. 하지만, 비록 이번 연구에서는 나이와 OPS 관계가 높지않았다

〈표 12〉 [SET 1]에 대한 정확도,재현율,정밀도, F1

2년치 데이터로 예측	SET 1							
	accuracy		recall		precision		f1	
	나이 포함		나이 포함		나이 포함		나이 포함	
	O	X	O	X	O	X	O	X
Nearest Neighbors	0.68	0.56	0.79	0.88	0.59	0.49	0.68	0.63
Linear SVM	0.42	0.42	1.00	1.00	0.42	0.42	0.59	0.59
RBF SVM	0.72	0.74	0.75	0.83	0.64	0.65	0.69	0.73
Gaussian Process	0.74	0.74	0.83	0.83	0.65	0.65	0.73	0.73
Decision Tree	0.68	0.77	0.83	0.83	0.59	0.69	0.69	0.75
Random Forest	0.63	0.77	0.54	0.83	0.57	0.69	0.55	0.75
Neural Net	0.72	0.70	0.79	0.92	0.63	0.59	0.70	0.72
AdaBoost	0.68	0.81	0.71	0.75	0.61	0.78	0.65	0.77
Naive Bayes	0.72	0.72	0.88	0.79	0.62	0.63	0.72	0.70
QDA	0.63	0.74	0.83	0.83	0.54	0.65	0.66	0.73
XGB	0.68	0.81	0.75	0.75	0.60	0.78	0.67	0.77

〈표 13〉 [SET 2]에 대한 정확도,재현율,정밀도, F1

2년치 데이터로 예측	SET 2							
	accuracy		recall		precision		f1	
	나이 포함		나이 포함		나이 포함		나이 포함	
	O	X	O	X	O	X	O	X
Nearest Neighbors	0.58	0.58	0.64	0.75	0.55	0.56	0.64	0.60
Linear SVM	0.49	0.49	1.00	1.00	0.49	0.49	0.66	0.66
RBF SVM	0.68	0.63	0.61	0.71	0.67	0.63	0.69	0.62
Gaussian Process	0.67	0.58	0.61	0.71	0.65	0.57	0.68	0.59
Decision Tree	0.68	0.61	0.57	0.71	0.67	0.62	0.69	0.59
Random Forest	0.63	0.61	0.57	0.75	0.60	0.62	0.67	0.59
Neural Net	0.65	0.61	0.61	0.79	0.61	0.61	0.69	0.61
AdaBoost	0.70	0.70	0.64	0.79	0.67	0.72	0.72	0.68
Naive Bayes	0.63	0.61	0.61	0.68	0.61	0.61	0.64	0.61
QDA	0.60	0.63	0.61	0.82	0.56	0.63	0.67	0.62
XGB	0.61	0.61	0.57	0.79	0.58	0.62	0.67	0.59

〈표 14〉 [SET 1]에 대한 변수 중요도 (7년치의 데이터로 OPS를 예측했을 경우)

입력값	순열 중요도						특성중요도					
	decision_tree		random_forest		xgb		decision_tree		random_forest		xgb	
	나이포함여부		나이포함여부		나이포함여부		나이포함여부		나이포함여부		나이포함여부	
	O	X	O	X	O	X	O	X	O	X	O	X
-7년 나이	0.054		0.038		0.073		0.107		0.085		0.034	
-7년 OPS	0.030	0.022	0.070	0.040	0.130	0.128	0.060	0.065	0.088	0.132	0.068	0.135
-6년 나이	0.026		0.042		0.048		0.034		0.117		0.091	
-6년 OPS	0.023	0.023	0.013	0.037	0.066	0.074	0.061	0.065	0.045	0.113	0.051	0.080
-5년 나이	0.021		0.050		0.070		0.058		0.038		0.078	
-5년 OPS	0.073	0.105	0.081	0.038	0.125	0.118	0.080	0.172	0.090	0.075	0.067	0.118
-4년 나이	0.000		0.023		0.004		0.000		0.029		0.059	
-4년 OPS	0.035	0.04	0.031	0.048	0.094	0.127	0.105	0.129	0.050	0.172	0.076	0.141
-3년 나이	0.007		0.061		0.019		0.026		0.099		0.073	
-3년 OPS	0.025	0.039	0.012	0.056	0.096	0.120	0.042	0.071	0.050	0.124	0.072	0.136
-2년 나이	0.000		0.048		0.010		0.000		0.058		0.054	
-2년 OPS	0.014	0.078	0.013	0.017	0.076	0.078	0.068	0.118	0.056	0.091	0.047	0.105
-1년 나이	0.000		0.012		0.057		0.000		0.077		0.082	
-1년 OPS	0.208	0.226	0.090	0.168	0.292	0.282	0.352	0.377	0.111	0.289	0.139	0.282

〈표 15〉 [SET 2]에 대한 변수 중요도 (7년치의 데이터로 OPS를 예측했을 경우)

입력값	순열 중요도						특성중요도					
	decision_tree		random_forest		xgb		decision_tree		random_forest		xgb	
	나이포함여부		나이포함여부		나이포함여부		나이포함여부		나이포함여부		나이포함여부	
	O	X	O	X	O	X	O	X	O	X	O	X
-7년 나이	0.029		0.033		0.077		0.083		0.084		0.046	
-7년 OPS	0.088	0.067	0.025	0.042	0.124	0.136	0.120	0.127	0.081	0.149	0.058	0.113
-6년 나이	0.017		0.011		0.066		0.043		0.093		0.052	
-6년 OPS	0.023	0.024	0.005	0.046	0.095	0.056	0.047	0.050	0.042	0.055	0.053	0.069
-5년 나이	0.000		0.025		0.018		0.000		0.048		0.077	
-5년 OPS	0.046	0.076	0.063	0.035	0.081	0.061	0.044	0.161	0.102	0.132	0.054	0.111
-4년 나이	0.000		0.024		0.001		0.000		0.049		0.024	
-4년 OPS	0.075	0.065	0.008	0.032	0.050	0.104	0.133	0.118	0.048	0.129	0.087	0.107
-3년 나이	0.000		0.008		0.021		0.000		0.047		0.060	
-3년 OPS	0.041	0.048	-0.001	0.066	0.088	0.068	0.050	0.053	0.044	0.175	0.056	0.131
-2년 나이	0.011		-0.000		0.007		0.023		0.100		0.077	
-2년 OPS	0.011	0.038	0.033	0.035	0.032	0.047	0.043	0.094	0.041	0.046	0.048	0.090
-1년 나이	0.040		0.032		0.017		0.054		0.061		0.111	
-1년 OPS	0.290	0.289	0.116	0.188	0.294	0.282	0.357	0.393	0.153	0.310	0.191	0.376

고 하더라도 선수의 나이는 연봉협상이나 팀 전략 구상을 할 때 고려해야 하는 중요한 항목이므로, 향후에는 기록으로 “야구 나이”를 예측할 수 있는 연구의 필요성이 제기된다.

이렇게 머신러닝 모델을 사용하여 타자의 OPS의 상승 및 하락을 예측할 수 있다면 선수 영입 및 다음해 전략 구상에 보다 효율적인 운영을 할 수 있을 것이다. 뿐만 아니라, 시즌 후에 머신러닝으로 예측한 OPS와 타자의 실제 OPS를 비교하는 서비스를 제공한다면, 관중들의 흥미도 더 높일 수 있을 것이다.

본 연구에서는 스탯티즈[11]에 나온 기록 그대로 실험에 적용하였고, 타자 성적의 외부 조건이라고 할 수 있는 그 해의 투수들의 기량, 스트라이크 존, 공의 반발 계수 등은 반영하지 못했다. 예를 들어, 평균 OPS가 .700인 리그에서 OPS .700을 기록한 것과 평균 OPS .800에서 OPS .700을 기록한 것을 동일하게 처리하였다. 이와 같은 외부 조건 들을 모델에 반영할 수 있다면, 보다 더 정확하게 최적의 OPS를 예측할 수 있을 것이다. 또, 투수력에 따라서 OPS가 달라질 수 있으므로, 각 시즌별로 투수력을 고려한 OPS를 산정한 뒤 실험을 하면 더 정확한 결과를 얻을 수 있을 것이다.

그리고, 야구에서는 포지션별로 체력 소모가 달라질 수 있는데, OPS 상승 및 하락 예측에 포지션도 고려된다면 더 정확한 예측이 될 수 있을 것이다.

이번 연구는 OPS를 11개의 범위로 나누고, 그것에 대한 상승 또는 하락을 예측한 것이다. 조금 더 OPS의 범위를 세분화 하여 상승 및 하락을 예측할 수 있다면, 실제 팀 전략 구상에 더 많이 도움을 줄 수 있을 것이다.

참 고 문 헌

[1] KBO 2020 공식 야구규칙 https://lgcxydabfbch3774324.cdn.ntruss.com/KBO_FILE/ebook/pdf/2020_야구규칙.pdf

[2] Blakeley B. McShane, Alexander Braunstein, James Piette and Shane T. Jensen, “A Bayesian Variable Selection Approach to Major League Baseball Hitting Metrics,” Journal of Quantitative Analysis in Sports, Vol. 7, No. 4, 2009.

[3] 조영석, 조영주, “한국프로야구에서 OPS와 득점에 관한 연구”, Journal of The Korean Data Analysis Society, Vol. 7, No. 1, pp. 221-231, 2005.

[4] 김혁주, “한국프로야구에서 출루능력과 장타력이 득점 생산성에 미치는 영향”, 한국데이터정보과학회지, Vol. 23, No. 6, pp. 1165-1174, 2012.

[5] 정진상, “빅 데이터 분석 기법을 이용한 한국프로야구 타자 평가 지표 개발”, 창원대학교 석사학위논문집, 2014.

[6] 정예린, “빅 데이터 분석과 투수 기량을 반영한 한국프로야구 타자 평가 모델”, 창원대학교, 2017.

[7] 김예형, “한국 프로야구에서 득점과 실점에 영향을 미치는 요인에 관한 통계적 연구”, 원광대학교, 2014.

[8] 문형우, “야구 경기에서 빅데이터 분석과 마르코프 연쇄를 이용한 득점 예측 모형”. 창원대학교 박사학위논문, 2014.

[9] 이승준, “데이터가 바꿀 한국 야구의 미래”, 한겨레, Oct. 27, 2019.

[10] 장원석, “예전의 명 감독은 잊어라, 데이터가 우승을 이끈다”, 동아비즈니스리뷰, Vol. 286, 2019.

[11] 스탯티즈, http://www.statiz.co.kr/

[12] 김민택, 구자환, 김응모, “하둡 및 스파크 기반 빅데이터 분석 플랫폼을 이용한 타자 OPS 예측”, 한국정보과학회 학술발표논문집 Vol. 2019, No.12.

[13] 박지훈, http://suxism.com/?page_id=3453
[14] 한정섭, 정다현, 김성준, “머신러닝을 활용한 빅데이터 분석을 통해 KBO 타자의 OPS 예측”, 차세대융합기술학회논문지 Vol. 6, No. 1, 2022.
[15] 홍종선, 신동식, “2017년 한국프로야구 타자력 예측모형 개발”, 한국데이터정보과학회지 Vol. 28, No. 3, 2017.
[16] 유진, 조선일보, 2022.02.04 https://www.chosun.com/sports/sports_photo/2022/02/04/WYAO6SJZOOPPAOFDSMN5A5TTBM/
[17] 벤저민 바우머, 앤드루 짐발리스트, 세이버메트릭스 레볼루션, 송민구 역, 한빛비즈, 2015.
[18] 나무위키, https://namu.wiki/w/OPS?rev=197
[19] 황규인, 동아일보, 2020-05-14 https://www.donga.com/news/Sports/article/all/20200514/101035162/1
[20] Mitchel Lichtman, 2009-12-21, https://tht.fangraphs.com/how-do-baseball-players-age-part-1/
[21] 이장택, “한국프로야구 타자 연봉의 결정요인”, 한국데이터정보과학회지, Vol. 30, No. 6, pp. 1375-138, 2019.
[22] 오상진, 스포탈코리아, 2018-05-16, https://sports.v.daum.net/v/20180516164422799
[23] 박성배, 이완영, 전홍권, “한국프로야구 타자 연봉 평가 척도에 영향을 미치는 경기력 변수 분석”, 한국사회체육학회지, Vol. 66, pp. 55 - 65, 2016.

저 자 소 개

신 동 윤(Dong Yun Shin)
·2020년 2월 : 강원대학교 컴퓨터과학과(박사과정 수료)
·2006년 9월~현재 : 클로잇(쌍용정보통신) 스포츠사업팀
·관심분야 : 빅데이터 활용, 클라우드 컴퓨팅, 딥러닝, 스포츠데이터 분석

김 진 호(Jinho Kim)
·1982년 2월 : 경북대학교 전자공학과(공학사)
·1984년 2월 : KAIST 전산학과(공학석사)
·1990년 2월 : KAIST 전산학과(공학박사)
·1990년 8월~현재 : 강원대학교 컴퓨터공학과 교수
·관심분야 : 대용량 빅데이터 저장 및 처리, 하둡/맵리듀스 분산/병렬처리 기술, 빅데이터 분석 기법, 데이터 마이닝, 클라우드 컴퓨팅, 데이터 웨어하우스, OLAP 다차원 분석, 데이터베이스 시스템 개발