2021-05

[스포츠 분석] 세이버 스탯을 이용한

한국 프로야구 선수의 연봉 데이터 분석

김 진 수

이 예 리

임 재 균

최 정 원

홍 기 현

글로벌아이티인재개발원

요 약 서 (초 록)

과 제 명	[스포츠 분석] 세이버 스탯을 이용한 한국 프로야구 선수의 연봉 데이터 분석
소 속	글로벌아이티인재개발원
프로젝트원	김진수, 이예리, 임재균, 최정원, 홍기현
프로젝트기간	2021. 05. 03. ~ 2021. 05. 28
Key Word	**야구경기분석, 머신러닝, 타자, 투수분석**

1. 프로젝트의 필요성 및 목적

o 한국 프로야구에 대한 전 국민적 인기와 관심

o 기존의 데이터 분석 프로젝트의 한계

- '클래식 스탯'을 사용한 데이터 분석 및 팀 관련 데이터만을 사용

o '세이버 스탯'을 바탕으로 한 객관적인 선수 연봉 분석

2. 프로젝트내용 및 범위

o 한국 프로야구 10개 팀별 타자와 투수의 3개년(2018~2020년) 데이터

3. 프로젝트방법

o 세이버 스탯 데이터 사용

o 머신러닝 지도 학습 중 회귀분석 방법 사용

4. 결론

o 투수 연봉 예측 모델이 타자 연봉 예측 모델보다 정확도가 높음

o 연봉에 대한 전관예우 등의 현실적인 한계로 인해 연봉 예측 모델의 값이 이상치로 나타나는 경우가 있음

o 종속 변수인 연봉에 가장 유의미한 영향을 주는 독립 변수는 타자와 투수 WAR임

목 차

<표 목차>

<그림목차>

I

서 론

I 서론

1. 프로젝트 필요성 및 목적

2019년 코로나바이러스 감염증(이하 코로나)의 발병 이전까지 프로야구의 경기당 평균 관중 수는 2015년 이후 지속해서 증가하고 있으며, 2019년 평균 관중 수는 10,280명을 기록하여 아시아 모든 리그 중 평균 관중 수 7위를 기록했다(e-나라지표, 2021). 실제로 설문조사 결과 '프로야구에 관심이 있다'는 답변을 한 국민이 41.2%를 차지할 정도로 프로야구는 국민의 폭넓은 인기와 관심을 독차지하고 있다(한국갤럽, 2021).

이러한 야구에 대한 국민의 관심은 최근 리그 개막에 맞추어 선수들의 리그 성적 데이터를 활용하는 '아웃 오브 더 파크 베이스볼(이하 OOTP 베이스볼)', '컴투스 프로야구 2021', '프로야구 매니저', '마구마구' 등의 PC와 모바일 게임에 관한 관심으로 이어졌다. 특히, 'OOTP 베이스볼' 게임은 다양한 국가들의 야구 데이터를 이용하여 지난 4년 3번이나 MLB 월드시리즈의 우승팀 예측에 성공한 바 있다. 사회과학의 게임 이론과 통계학적 방법론을 도입하여 야구의 객관적인 지식을 추구하는 '세이버매트릭스(SABERmetrics)'와 더불어 빅데이터의 발전에 힘입어 선수, 경기, 팀의 빅데이터를 이용하여 경기 결과를 예측하는 '데이터 야구'가 보편화되고 있다.

이전의 야구에 대한 데이터 분석은 로지스틱 회귀분석과 의사결정나무모형(CHAID 기법) 등을 사용하여 프로야구 승·패 예측모형에 관한 프로젝트를 수행하였고(Kim, 2001), ID3와 통계적 방법, 역전파 알고리즘을 사용하여 승패 예측시스템을 구축하는 프로젝트를 수행하였다(Hong et al, 2003). 그러나 기존의 데이터 분석 방법론의 대부분은 타율, 팀 평균 자책점 등의 팀 관련 데이터를 사용하거나 클래식 스탯을 사용하여 선수 개개인에게 초점을 맞추지 않았다는 한계가 존재한다. 따라서 본 프로젝트에서는 세이버 스탯을 이용하여 한국 프로야구 타자와 투수 개개인의 역량 및 기존 연봉 데이터를 기반으로 미래 연봉을 예측해보고자 한다.

빅데이터 분석을 통해 프로야구의 선수의 연봉을 예측하는 것은 다음과 같은 의의가 있다. 첫째, 데이터 분석을 통해 선수의 기량을 평가하여 객관적 수치에 근거하여 연봉 협상을 하는 데 도움을 줄 수 있다. 둘째, 시즌 도중의 결과 예측을 통해 스포츠 토토와 프로토 등과 같은 체육 진흥 투표권의 이용자들에게 투자 정보를 제공할 수 있고(Koo et al., 2009; Oda-chowski and Grekow, 2013), 프로야구 게임 사용자와 야구팀의 팬들에게 즐거움을 고취할 수 있다.

2. 프로젝트 방법 및 프로젝트 추진 절차

가. 클래식 스탯과 세이버 스탯

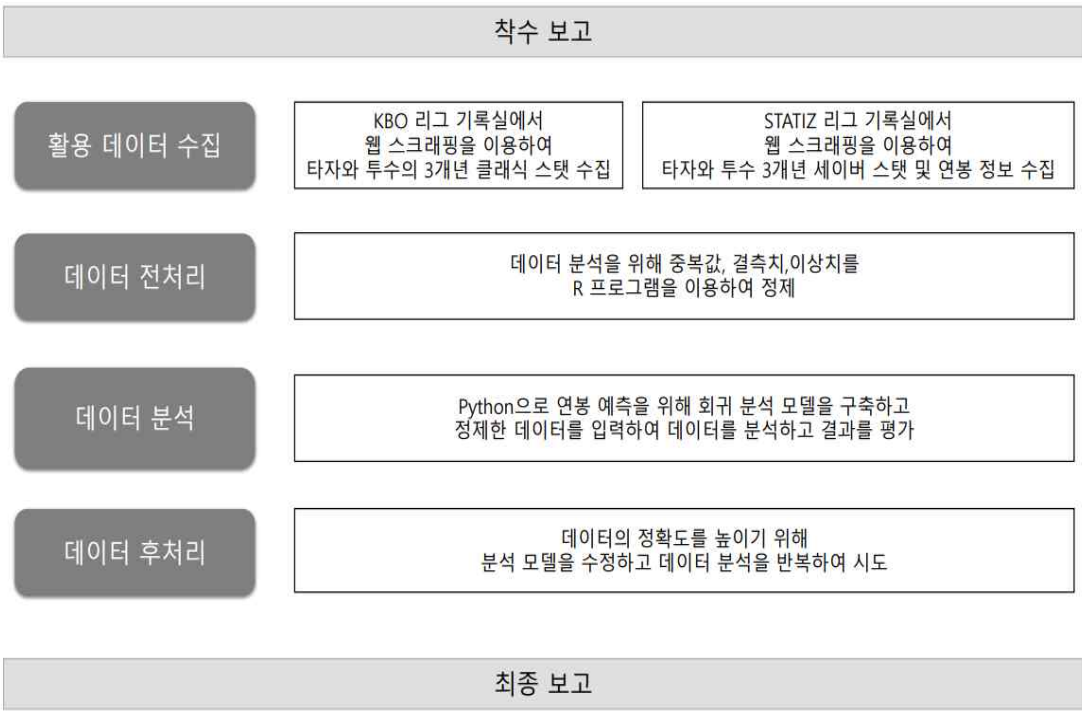
야구는 한 경기 당 약 1,000개의 데이터가 산출될 정도로 데이터가 많은 스포츠 종목이다. 기존의 데이터 분석은 수치화하기 쉬운 클래식 스탯(Classic Stats)을 사용하여 이해하기 쉽고 가독성이 높았으나, 오류가 많고 예측이 정확하지 않다는 한계가 존재한다. 따라서 본 프로젝트는 선수들의 능력을 자세하고 정확하게 측정하기 위해 기존 클래식 스탯에 각종 과학적 장비 및 계산식을 적용한 세이버 스탯(Saber Stats)의 3개년 데이터를 바탕으로 2020년 선수들의 연봉을 예측해보고자 한다.

나. 머신러닝(Machine Learning) 분석 방법론

미래 연봉을 예측하기 위해서는 기계 스스로 데이터를 기반으로 패턴을 학습하고 결과를 예측하는 머신러닝 방법론을 사용해야 한다. 머신러닝은 알고리즘에 주입하는 훈련 데이터인 레이블(Label)의 포함 여부에 따라 지도 학습(supervised learning)과 비지도 학습(unsupervised learning)으로 나뉜다. 본 프로젝트에서는 2017년, 2018년, 2019년 타자와 투수의 세이버 스탯 데이터와 2017년, 2018년 연봉 데이터를 독립 변수로 하여, 2020년 연봉을 종속 변수로 두어 예측을 실시하고자 한다. 특히, 본 프로젝트에서 사용할 분석 방법은 여러 개의 독립 변수와 하나의 종속 변수 간의 관계를 모형화하여 예측하고, 모든 변수가 연속형인 점을 비추어 보았을 때, 지도 학습 중 회귀 분석을 시행하는 것이 제일 적합하다고 판단하였다.

다. 프로젝트 추진 절차

2020년 선수들의 연봉 예측과 나아가 연봉 예측 모델 구축을 위해 본 프로젝트는 [그림 Ⅰ-1] <프로젝트 추진 절차>에 제시한 바와 같이 진행하였다. 프로젝트의 효율적인 진행을 위해 프로젝트 간 분야를 분장하여 단계적 절차 없이 포괄적으로 수행하고 프로젝트결과를 종합하였으며, 예측 모델을 제시하였다.



[그림 Ⅰ- 1] 프로젝트 추진 절차

Ⅱ

데이터 수집

1. 데이터 분석 범위
2. 데이터 수집
   가. KBO 리그 기록실
   나. STATIZ 리그 기록실

Ⅱ 데이터 수집

1. 데이터 분석 범위

실제 연봉 데이터와 비교를 위해 '2020년 연봉' 을 종속 변수로 데이터 분석을 시행하기 때문에, 종속 변수에 영향을 줄 수 있는 '2018년, 2019년, 2020년의 세이버스탯으로 평가한 선수의 역량' 및 '2018년, 2019년의 연봉 데이터' 를 독립 변수를 선정하여 분석을 시행한다. 특히, 야구에서는 타자와 투수의 역량을 평가하는 세이버스탯의 변수가 다르므로, 선수는 타자와 투수로 이원화하여 분석을 시행한다.

한편, 1개년 데이터만을 분석할 경우, 부상 등으로 인해 일반적이지 않은 영향을 받을 수 있기 때문에 3개년의 데이터를 수집하여 일관성을 확보하고자 한다. 세이버스탯을 계산하기 위한 클래식 스탯 데이터는 KBO 공식 홈페이지 기록실[1]에서 수집하였으며, 세이버 스탯 데이터와 연봉 데이터는 기록실[2]을 이용하였다. 본 프로젝트에서 사용한 세이버 스탯 변수는 다음과 같다.

1) https://www.koreabaseball.com/Record/Main.aspx (21.05.24 방문)
2) http://www.statiz.co.kr/main.php (21.05.24 방문)

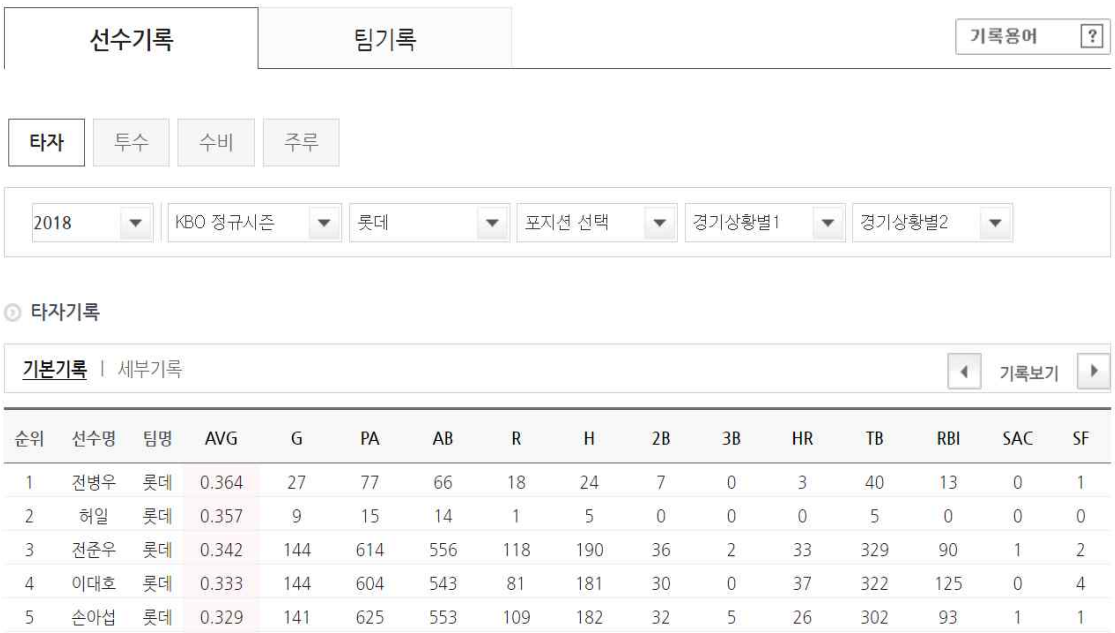
변수		설명
Batter (타자)	WAR	**Wins Above Replacement** 대체선수와 비교했을 때, 얼마나 많은 승리에 기여했는가를 나타내는 수치
	wRC	**Weighted Runs Created** wOBA에 기반을 둔 타격으로 얻어낸 득점기여도
		$(\frac{wOBA - 리그wOBA}{wOBA스케일} + \frac{리그득점}{타석}) * 타석$
	wRAA	**Weighted Runs Above Average** 평균적인 선수와 비교해서 타격으로 얻어낸 득점기여도
		$\frac{wOBA - 리그\ wOBA}{wOBA\ 스케일} * 타석$
	wOBA	**weighted On Base Average** 출루 이벤트별 실제 득점 가치에 비례한 가중치를 부여한 출루율
Pitcher (투수)	FIP	**Fielding Independent Pitching** ERA의 단점을 보완한 스탯으로, 전적으로 투수에게 책임이 있다고 생각되는 기록들만을 추린 평균자책점의 형태
	LOB%	**Left On Base Percentage** 출루 된 주자 중 득점하지 않은 비율을 나타내는 수치
		$\frac{안타 + 볼넷 + 사구 - 실점}{안타 + 볼넷 + 사구 - (1.4 * 피홈런)}$
	BABIP	**Batting Average on Balls in Play** 인플레이 타구의 안타 비율 혹은 피안타 비율

<표 Ⅱ- 1> 활용 세이버 매트릭스 변수

2. 데이터 수집

가. KBO 리그 기록실

KBO 리그 기록실에서 타자의 H(안타), HR(홈런), RBI(타점)의 클래식 스탯 데이터, 투수의 ERA(평균자책점), IP(이닝), WHIP(이닝당 출루허용률)의 클래식 스탯 데이터를 웹 스크래핑(Web Scraping, 웹 사이트에서 원하는 데이터를 가져오는 행위)을 통해 수집하고, 엑셀을 이용하여 정리한다.


선수기록 | 팀기록 | 기록용어

타자 | 투수 | 수비 | 주루

2018 | KBO 정규시즌 | 롯데 | 포지션 선택 | 경기상황별1 | 경기상황별2

타자기록

기본기록 | 세부기록 | 기록보기

순위	선수명	팀명	AVG	G	PA	AB	R	H	2B	3B	HR	TB	RBI	SAC	SF
1	전병우	롯데	0.364	27	77	66	18	24	7	0	3	40	13	0	1
2	허일	롯데	0.357	9	15	14	1	5	0	0	0	5	0	0	0
3	전준우	롯데	0.342	144	614	556	118	190	36	2	33	329	90	1	2
4	이대호	롯데	0.333	144	604	543	81	181	30	0	37	322	125	0	4
5	손아섭	롯데	0.329	141	625	553	109	182	32	5	26	302	93	1	1

[그림 Ⅱ-1] KBO 기록실 이미지

나. STATIZ 기록실

STATIZ 기록실에서 수치화되어 있는 세이버 스탯과 선수별 연봉을 웹 스크래핑을 통해 수집하고, 엑셀을 이용하여 정리한다.

시즌기록실

종합 타격 투구 수비

2021 2020 시작 끝 두산 포지션 정규 규정 상황 옵션

[자동 : 전체]

기본 확장 가치 클러치 타석 타구1 타구2 파워 팀배팅1 팀배팅2 도루 주루 구종가치 구종구사

순	이름	팀	생산력+ wRC+	타석	HR%	BB%	K%	BB/K	IsoP	IsoD	BABIP	Spd	PSN	타격 생산력 wOBA	wRC	$wRC_{/27}$	wRAA	타격 생산력 (파크팩터 조정) wOBA	wRC	$wRC_{/27}$	wRAA	wRC+
1	신성현	20 두	202.0	7	0.00	42.9	42.9	1.00	.250	.321	1.000	3.9		.498	1.9	16.69	0.9	.499	1.9	16.71	0.9	202.0
2	류지혁	20 두 3B	180.6	31	3.23	3.2	16.1	0.20	.208	.050	.450	3.0	0.00	.457	7.1	10.09	3.0	.466	7.3	10.44	3.3	180.6
3	페르난데스	20 두 DH	146.5	668	3.14	8.7	6.3	1.38	.157	.065	.333	1.7	0.00	.406	123.6	7.85	35.9	.414	128.5	8.16	40.8	146.5
4	오재일	20 두 1B	138.8	534	3.00	11.4	17.2	0.66	.170	.077	.359	2.3	3.56	.394	93.3	7.43	23.2	.403	97.3	7.75	27.2	138.8
5	김재환	20 두 LF	135.4	613	4.89	14.9	25.1	0.59	.229	.108	.318	3.4	10.00	.385	102.3	6.94	21.9	.397	108.9	7.39	28.5	135.4
6	박건우	20 두 RF	126.4	551	2.54	7.6	11.8	0.65	.168	.065	.323	4.7	10.18	.375	87.5	6.56	15.1	.384	91.5	6.86	19.1	126.4
7	최주환	20 두 2B	123.7	573	2.79	8.4	11.5	0.73	.167	.060	.319	3.8	3.56	.372	89.2	6.46	14.0	.380	93.0	6.74	17.8	123.7
8	허경민	20 두 3B	122.7	487	1.44	7.2	5.8	1.25	.110	.050	.337	4.5	9.33	.373	76.5	6.49	12.5	.378	78.5	6.66	14.5	122.7
9	정수빈	20 두 CF	104.5	559	0.89	9.8	10.0	0.98	.098	.070	.325	6.5	7.50	.348	75.5	5.48	2.1	.350	76.7	5.57	3.3	104.5
10	김재호	20 두 SS	95.3	457	0.44	9.9	9.9	1.00	.057	.073	.318	3.1	3.00	.334	56.4	4.89	-3.6	.336	57.2	4.96	-2.8	95.3

[그림 II-2] STATIZ 기록실 이미지

Ⅲ

데이터 전처리

III

R을 이용한 데이터 전처리

1. 이상치와 결측치 제거

데이터 전처리란 수집 데이터 중 분석 결과나 모델 성능에 악영향을 미칠 수 있는 값을 가공하는 작업을 의미한다. 즉, 원활한 데이터 분석을 위해 데이터 자체 혹은 수집 과정에서 발생한 오류를 데이터 전처리 과정을 통해 정제하는 것이다.

본 프로젝트에서는 엑셀로 정렬된 원본 데이터에서 이상치와 결측치를 제거하는 과정을 진행하였다. 사용 프로그래밍 언어는 ‘R’이며 구현 환경은 ‘R Studio’를 사용하였다.

가. 이상치 제거

이상치(Outlier)란 사람의 나이에 대한 데이터에 음의 정수 값이 삽입된 것처럼 정상 범주에서 벗어난 값을 의미한다. 데이터가 포함된 채 데이터 분석이 진행될 경우 결괏값에 영향을 미친다. 따라서 데이터 정제 과정에서 이상치 제거는 필수적이다.

나. 결측치 제거

결측치(Missing Value)란 관측 대상 변수에 어떠한 값도 들어가지 않은 상태를 의미한다. 즉, 결측치란 손실 데이터를 일컫는다. 결측치를 유지한 채 데이터 분석 과정을 거칠 경우, 분석 결론에 영향을 미치기 때문에 사전 제거 작업이 필요하다.

본 프로젝트에서는 3개년 각각의 세이버 스탯을 평균으로 도출하여 독립변수로 사용하였다. 이때, 해외 진출 등으로 3개년 세이버 스탯의 평균을 구할 수 없는 선수는 결측치로 처리하여 분석에서 제거하였다.

2. 필요 데이터 추출

세이버 스탯은 기존의 클래식 스탯을 수학적 수식을 통해 도출하기 때문에 세이버 스탯을 계산하여 변수를 생성하고, 필요하지 않은 변수를 제거하여 데이터 분석에 필요한 데이터를 추출하는 과정을 거친다.

가. LOB% 변수 생성

세이버 스탯 변수 중 투수의 LOB%(Left on Base Percentage, 잔루처리율)[3]에 대한 수치가 수집 데이터에 존재하지 않기 때문에, 수집한 클래식 스탯 데이터를 바탕으로 공식을 통해 LOB% 변수 값을 계산한다.

나. 불필요 데이터 제거

타자와 투수로 이원화한 데이터에서 각각 불필요한 데이터를 제거하고, 데이터 분석에 필요한 세이버 스탯 변수만을 정제한다. 이를 통해 데이터 분석 시 시각적으로 표현할 수 있으며, 불필요한 데이터를 파악하는 데 사용되는 시간을 줄여 구현 환경을 용이하게 할 수 있다.

이상의 절차를 통해 143명의 타자 데이터와 102명의 투수 데이터를 추출하였다. 이를 CSV 파일로 저장하여, 이후 데이터 분석 과정은 정제된 CSV 파일을 이용한다.

3) $\frac{\text{안타}+\text{볼넷}+\text{사구}-\text{실점}}{\text{안타}+\text{볼넷}+\text{사구}-(1.4*\text{피홈런})}$

IV

데이터 분석

1. Python을 이용한 회귀분석

가. 타자(Batter)

1) 데이터 탐색
2) 데이터 분석
3) 회귀분석 변수 및 모델 평가
4) 분석 결과 시각화

나. 투수(Pitcher)

1) 데이터 탐색
2) 데이터 분석
3) 회귀분석 변수 및 모델 평가
4) 분석 결과 시각화

Ⅳ 데이터 분석

1. Python을 이용한 회귀분석

‘Python’으로 ‘Jupyter Notebook’에서 분석을 시행한다. 분석을 용이하게 하기 위해 ‘Pandas’, ‘Numpy’, ‘Matplotlib’의 3개의 라이브러리를 사용하였다.

```
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
```

[그림 Ⅳ-1] Python 라이브러리 사용

정제한 데이터를 데이터 분석 구현 환경에 가져온다. 한글을 Python에서 구현하기 위해 CP949 코드를 이용하여 인코딩하고, 투수와 타자는 각각 pitcher, batter 변수명으로 저장한다.

```
pitcher = pd.read_csv( "AllPitcher.csv", encoding = 'CP949')
batter = pd.read_csv("AllBatter.csv", encoding = 'CP949')
```

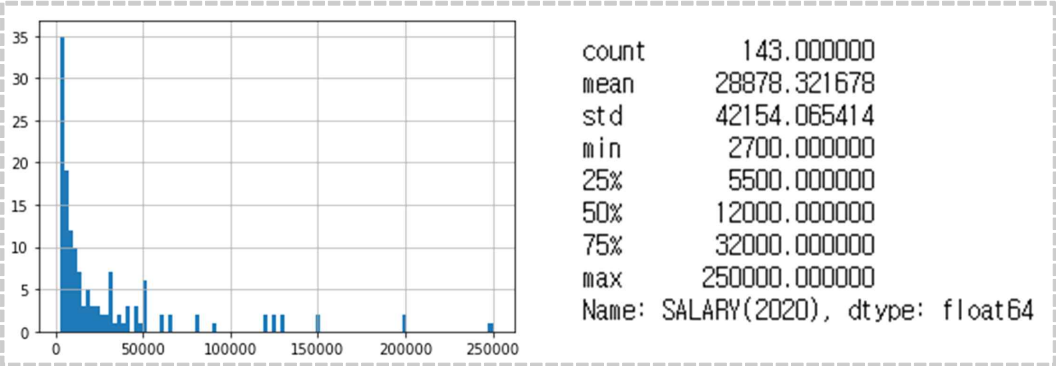
[그림 Ⅳ-2] CSV 파일 불러오기

각 선수의 2020년 연봉을 예측하는 데 필요한 변수(Feature)를 데이터 전처리 과정에서 ‘R’을 통해 정제했다. 따라서 데이터 분석 단계에서는 추가 데이터 처리 없이 분석을 시행한다.

가. 타자(Batter)

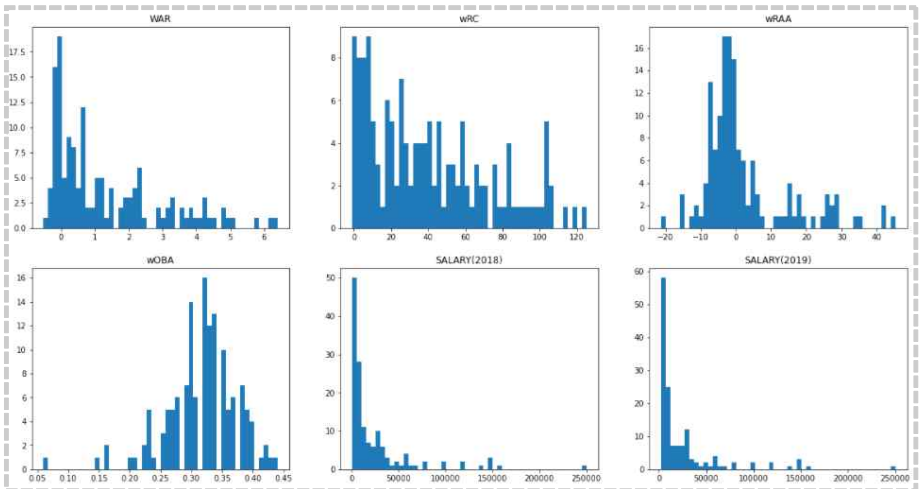
1) 데이터 탐색

예측할 대상인 종속 변수에 대한 데이터 분포를 파악하기 위해 실제 2020년 연봉 데이터를 확인한다.



[그림 Ⅳ-3] 타자의 종속 변수 데이터

먼저 회귀분석에 사용할 변수를 출력한다. 독립 변수는 WAR(선수 대비 승리 기여도), wRC(wOBA 기반 득점 생산), wRAA(리그 평균 대비 득점), wOBA(출루율 스케일), SALARY(2018)(2018년 연봉), SALARY(2019)(2019년 연봉) 이다.



[그림 Ⅳ-4] 타자 독립 변수 그래프

각 독립 변수의 값의 편차가 크기 때문에 큰 값을 지닌 독립 변수에 의해 종속 변수가 더 큰 영향을 받을 수 있다. 이를 해결하기 위해 z - 표준화 방법[4]을 이용한 피처 스케일링(Feature Scaling, 각 변수의 단위를 0 ~ 1 사이, 혹은 상대적 값을 표현할 수 있는 수치로 변경하는 것)을 한다.

	NAME	WAR	wRC	wRAA	wOBA	SALARY(2018)	SALARY(2019)	y
0	강경학	-0.376215	-0.401207	-0.355267	0.056818	-0.503903	-0.505972	7800
1	강민호	0.596649	0.480522	-0.115339	0.402564	1.950849	1.950861	125000
2	강백호	1.575711	1.859950	2.188631	1.439800	-0.584686	-0.586823	21000
3	강진성	-0.438181	-0.413495	-0.134237	-0.116054	-0.576868	-0.578999	3800
4	고종욱	-0.394805	0.170227	-0.451403	0.056818	-0.368397	-0.370351	17000

[그림 Ⅳ-5] 타자 피처 스케일링 후 변숫값

2) 데이터 분석

회귀분석을 시행하기 위해 필요한 라이브러리인 sklearn과 수학적 계산을 용이하게 하기 위해 math 라이브러리를 불러온다.

```
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from math import sqrt
```

[그림 Ⅳ-6] 회귀분석을 위한 Python 라이브러리

데이터셋을 학습 데이터(Train Data, 회귀분석 모델 구축을 위해 학습시킬 데이터셋)와 검증 데이터(Test Data, 회귀분석 모델의 정확도를 평가하기 위한 데이터셋)로 분리하고, 데이터의 수가 많지 않다는 한계를 반영하여 학습 데이터를 0.5로 설정하여 모델을 학습시켜 회귀 계수를 도출한다. 도출된 회귀 계수는 다음과 같다.

4) z - 표준화를 위한 계산식은 다음과 같다. $z = \frac{((x) - (x\text{의 평균})}{x\text{의 표준편차}}$

WAR	-10450.95388302	wRC	47773.55990713
wRAA	18671.74691477	wOBA	1809.21296527
SALARY(2018)	1583.25723169	SALARY(2019)	-12090.15447108

3) 회귀분석 변수 및 모델 평가

- 회귀분석 변수 평가

OLS Regression Results

Dep. Variable:	y	R-squared:	0.900
Model:	OLS	Adj. R-squared:	0.890
Method:	Least Squares	F-statistic:	95.74
Date:	Tue, 01 Jun 2021	Prob (F-statistic):	4.94e-30
Time:	16:26:38	Log-Likelihood:	-787.16
No. Observations:	71	AIC:	1588.
Df Residuals:	64	BIC:	1604.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.016e+04	1991.822	15.144	0.000	2.62e+04	3.41e+04
SALARY(2018)	-1.045e+04	7.24e+04	-0.144	0.886	-1.55e+05	1.34e+05
SALARY(2019)	4.777e+04	7.24e+04	0.660	0.512	-9.69e+04	1.92e+05
WAR	1.867e+04	8036.644	2.323	0.023	2616.705	3.47e+04
wOBA	1809.2130	3976.419	0.455	0.651	-6134.597	9753.023
wRAA	1583.2572	4428.036	0.358	0.722	-7262.761	1.04e+04
wRC	-1.209e+04	7969.576	-1.517	0.134	-2.8e+04	3830.904

Omnibus:	64.511	Durbin-Watson:	1.915
Prob(Omnibus):	0.000	Jarque-Bera (JB):	754.552
Skew:	2.324	Prob(JB):	1.42e-164
Kurtosis:	18.279	Cond. No.	120.

[그림 Ⅳ-7] 타자의 회귀분석 변수 평가

어떤 변수가 유의미한지 알아보기 위해 도출한 예측 모델의 변수를 평가한다. 회귀 분석의 예측도를 평가하는 지표인 결정계수(R-squared)는 모두 1에 근접할수록 분석의 예측도가 높다고 간주할 수 있다. [그림 Ⅳ-6]에서 주지하듯 결정계수(R-squared)는 0.900이며, 수정 결정 계수(Adj. R-squared)는 0.89으로, 분석의 예측도가 높다고 간주할 수 있다.

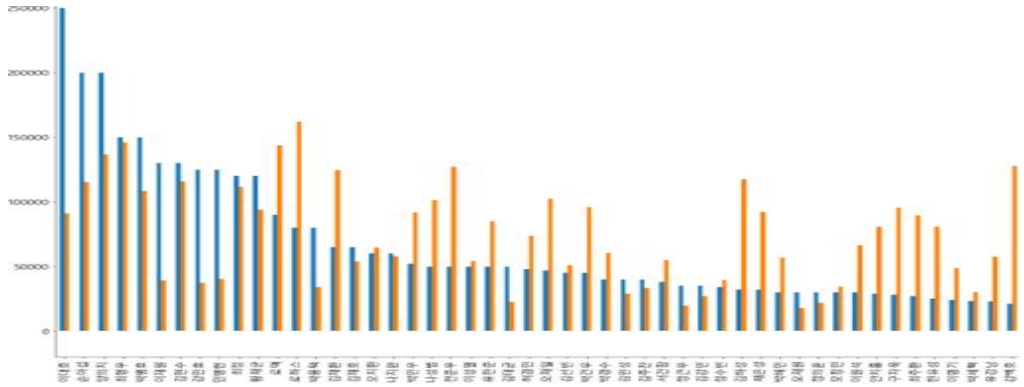
한편, [그림 Ⅳ-6]에서 P > |t| 수치는 각 변수의 검정 통계량이 얼마나 유의미한지를 나타내며, 유의 수준인 p-value(P > |t| 를 의미)의 값이 0.05 이하면 모델의 변수가 영향력 있는 변수임을 알 수 있다. WAR의 P > |t| 값이 0.023으로, 회귀분석에서 가장 영향력 있는 변수임을 알 수 있다. 또한, F-통계량(F-Statistic)에 따르면, 유의 수준인 p-value(Prob(F-Statistic)를 의미)의 값이 0.05 이하면 모델의 예측이 유의미하다. 타자 회귀분석 모델의 p-value 값이 4.94e-30 임으로 본 모델은 유의미하다고 간주할 수 있다.

- 회귀분석 예측모델 평가

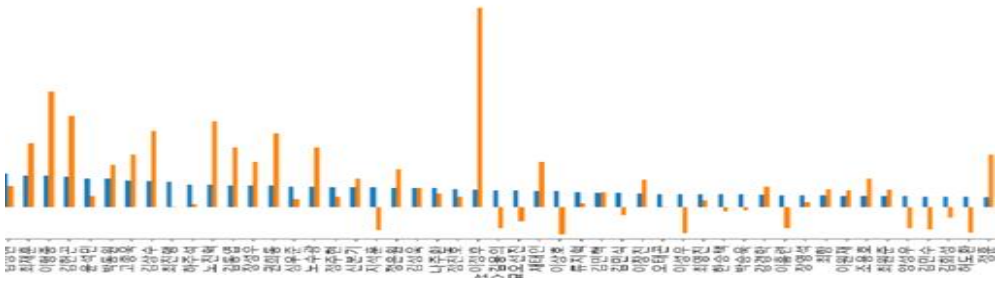
예측 모델을 평가하는 방법으로는 학습 데이터와 검증 데이터의 수정 결정 계수(R2 score)를 측정하는 방법과 둘째 평균 제곱근 편차(RMSE score)를 도출하여 방법이 있다. 그러나 평균 제곱근 편차에 대한 기준이 명확하지 않기 때문에, 본 프로젝트에서는 전자를 통해 회귀분석의 예측 모델을 평가하고자 한다. 두 데이터 간의 수정 결정 계수의 값이 차이가 작을수록 예측 모델의 정확도가 높다고 판단할 수 있다. 학습 데이터의 수정 결정 계수는 0.89975이며, 검증 데이터의 수정 결정 계수는 0.64027로, 모형의 예측도가 정확하다고 평가할 수 있다.

4) 분석 결과 시각화

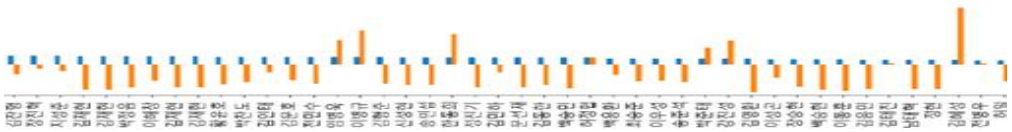
회귀분석 모델을 통해 예측한 2020년 타자의 연봉은 다음과 같다.



[그림 Ⅳ-8] 2020년 타자의 실제 연봉과 예측 연봉 그래프 <1>



[그림 Ⅳ-9] 2020년 타자의 실제 연봉과 예측 연봉 그래프 <2>



[그림 Ⅳ-9] 2020년 타자의 실제 연봉과 예측 연봉 그래프 <3>

한국 프로야구 선수의 연봉 데이터 분석

	선수명	실제연봉(2020)	예측연봉(2020)
0	이대호	250000	71525.963563
1	손아섭	200000	111383.536841
2	양의지	200000	148922.122637
3	최형우	150000	150257.280529
4	박병호	150000	104006.751844
5	이재원	130000	37223.852877
6	김현수	130000	114250.947886
7	강민호	125000	24962.361015
8	민병헌	125000	29620.193122
9	최정	120000	110521.897380
10	황재균	120000	88761.499320
11	로맥	90000	163581.448951
12	로하스	80000	177389.123077
13	박용택	80000	28867.306695
14	김재환	65000	136558.348582
15	김재호	65000	49377.944573
16	오지환	60000	64809.739118
17	나지완	60000	60127.743010
18	박민우	52000	100423.802324
19	나성범	50000	114324.634395
20	전준우	50000	145187.190262
21	이성열	50000	61864.153540
22	유한준	50000	92580.231785
23	김태균	50000	2770.890923
24	허경민	48000	79567.970370
25	오재일	47000	116230.040812
26	김선빈	45000	53464.949471
27	박건우	45000	104316.215087
28	박경수	40000	65831.342503
29	김민성	40000	25788.584268

	선수명	실제연봉(2020)	예측연봉(2020)
31	서건창	38000	59483.940108
32	정근우	35000	11975.484412
33	김강민	35000	21532.313089
34	정수빈	34000	40236.659244
35	김하성	32000	126860.005656
36	채은성	32000	105600.271427
37	박해민	30000	58025.874807
38	오재원	30000	10144.093110
39	정의윤	30000	22663.025310
40	모창민	30000	38871.240046
41	이원석	30000	72751.677634
42	안치홍	29000	88435.875726
43	구자욱	28000	108524.993214
44	최주환	27000	101740.798210
45	한유섬	25000	94755.181735
46	이명기	24000	52780.503915
47	박세혁	23200	30070.909783
48	유강남	23000	59648.966089
49	강백호	21000	152192.740183
50	이지영	21000	14391.047632
51	김성현	21000	8669.530949
52	최재훈	20000	42399.168173
53	이형종	20000	83424.460829
54	김헌곤	19000	66671.639426
55	윤석민	18000	3335.901700
56	박동원	18000	26249.692084
57	고종욱	17000	37991.546593
58	김상수	16500	50407.891697
59	최진행	16000	-1275.115308

[그림 Ⅳ-11] 타자의 실제 연봉과 예측 연봉 표

나. 투수(Pitcher)

1) 데이터 탐색

2020년 투수의 연봉에 대한 회귀분석 또한 타자의 경우와 동일한 절차를 통해 진행되었다. 독립 변수는 WAR(선수 대비 승리 기여도), FIP(수비 무관 추정 평균 자책점), LOB%(잔루율), BABIP(인플레이된 타구의 타율), SALARY(2018년 연봉), SALARY(2019)(2019년 연봉)이다. 투수의 경우도 각 변수를 피처 스케일링을 통해 정규화한다.

	NAME	WAR	FIP	LOB	BABIP	SALARY(2018)	SALARY(2019)	y
0	강동연	-0.683585	0.874631	0.469024	-0.421414	-0.433687	-0.434398	3400
1	강윤구	-0.651038	-0.175467	-0.257666	0.442591	-0.246190	-0.247566	15500
2	고우석	0.911198	-0.735008	0.469024	-0.637416	-0.394214	-0.395065	22000
3	구승민	0.466395	-0.428410	0.590139	-0.421414	-0.413950	-0.414732	8000
4	구창모	2.614471	-0.865312	1.437944	-0.637416	-0.338294	-0.339343	18000

[그림 Ⅳ-12] 투수 피처 스케일링 후 변숫값

2) 데이터 분석

데이터 수가 많지 않다는 한계를 반영하여 학습 데이터를 0.5로 설정하고 회귀분석 모델을 학습하여 회귀 계수를 도출한다. 도출된 회귀 계수는 다음과 같다.

WAR	128.50705397	FIP	906.17560653
LOB%	-486.0567842	BABIP	-2257.07667265
SALARY(2018)	28436.78053828	SALARY(2019)	7430.19818042

3) 회귀분석 변수 및 모델 평가

- 회귀분석 변수 평가

어떤 변수가 가장 유의미한지 알아보기 위해 도출한 예측 모델의 변수를 평가한다. 결정 계수(R-squared)는 0.905이며, 수정 결정 계수(Adj. R-squared)는 0.892으로 타자 예측 모델보다 근소한 차이로 투수 예측 모델이 더 정확도가 높다는 것을 알 수 있다. 또한 WAR의 유의 수준이 0.000으로 유의미한 변수라고 할 수 있다. [그림 Ⅳ-11] 투수의 회귀분석 변수 평가에서도 '2019년 연봉' 다음으로 WAR이 영향력이 큰 변수임을 알 수 있다.

Dep. Variable:	y	**R-squared:**	0.905
Model:	OLS	**Adj. R-squared:**	0.892
Method:	Least Squares	**F-statistic:**	69.84
Date:	Thu, 03 Jun 2021	**Prob (F-statistic):**	7.42e-21
Time:	11:47:14	**Log-Likelihood:**	-508.13
No. Observations:	51	**AIC:**	1030.
Df Residuals:	44	**BIC:**	1044.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.797e+04	828.903	21.676	0.000	1.63e+04	1.96e+04
BABIP	128.5071	1021.817	0.126	0.900	-1930.829	2187.843
FIP	906.1756	1059.712	0.855	0.397	-1229.534	3041.885
LOB	-486.0568	1044.093	-0.466	0.644	-2590.288	1618.174
SALARY(2018)	-2257.0767	8.7e+04	-0.026	0.979	-1.78e+05	1.73e+05
SALARY(2019)	2.844e+04	8.73e+04	0.326	0.746	-1.48e+05	2.04e+05
WAR	7430.1982	1264.034	5.878	0.000	4882.705	9977.691

Omnibus:	14.873	**Durbin-Watson:**	2.076
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	23.594
Skew:	-0.887	**Prob(JB):**	7.53e-06
Kurtosis:	5.821	**Cond. No.**	265.

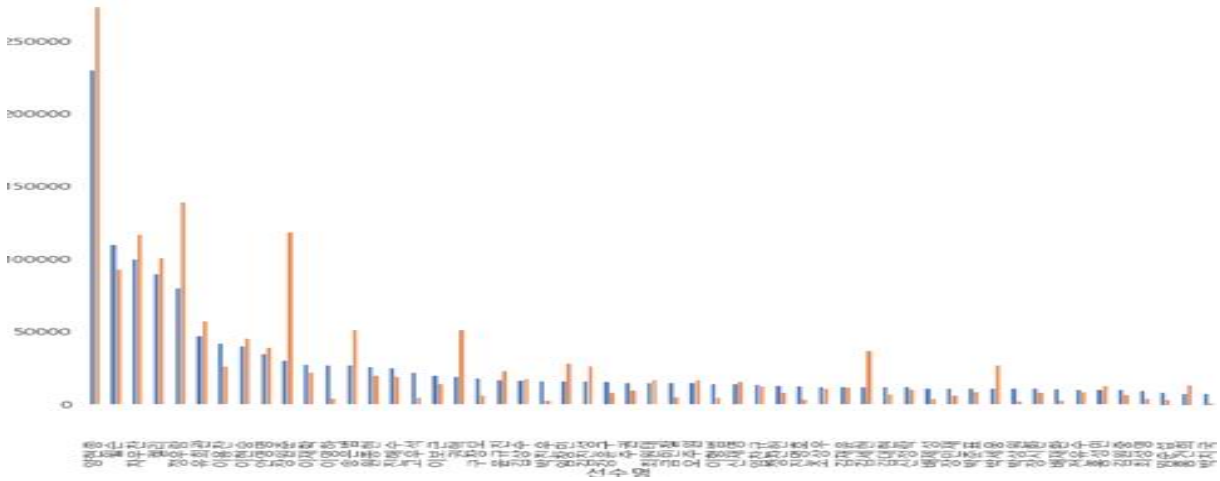
[그림 Ⅳ-13] 투수의 회귀분석 변수 평가

- 회귀분석 예측모델 평가

투수의 경우에서도 타자와 마찬가지로 수정 결정 계수(R2 score)를 측정하는 방법을 사용하여 예측 모델을 평가한다. 학습 데이터의 수정 결정 계수는 0.90497이며, 검증 데이터의 수정 결정 계수는 0.89872로 두 계수 간의 차이가 크지 않으므로 모형의 예측도가 정확하다고 평가할 수 있다.

4) 분석 결과 시각화

회귀분석 모델을 통해 예측한 2020년 투수의 연봉은 다음과 같다.



[그림 Ⅳ-14] 2020년 투수의 실제 연봉과 예측 연봉 그래프<1>

[그림 Ⅳ-14] 2020년 투수의 실제 연봉과 예측 연봉 그래프<2>

	선수명	실제연봉(2020)	예측연봉(2020)
0	양현종	230000	462650.432097
1	윌슨	110000	139095.041496
2	차우찬	100000	176411.532618
3	헥리	90000	154927.335534
4	정우람	80000	208831.727417
5	유희관	47000	74818.453766
6	이용찬	42000	24075.600703
7	이현승	40000	62285.489220
8	안영명	35000	40657.522830
9	장원준	30000	177102.850918
10	이재학	27500	27101.232342
11	이영하	27000	-2373.790289
12	송은범	27000	69063.425299
13	원종현	26000	14745.042097
14	진해수	25000	4207.397577
15	고우석	22000	7979.132662
16	이보근	20000	1660.808320
17	권혁	19000	63371.264424
18	구창모	18000	7678.127959
19	윤규진	17000	9125.147722
20	김상수	16500	15491.997587
21	박진우	16000	119.924966
22	임창민	16000	34163.752835
23	김진성	16000	35331.303679
24	강은구	15500	-8192.298853
25	주권	15000	22102.012232
26	최원태	15000	17398.409843
27	금민철	15000	-42382.442892
28	오주원	15000	19462.160583
29	이형범	14200	4923.684874
31	임찬규	13500	7264.673731
32	정찬헌	13000	-9677.098377
33	진명호	12500	2534.390625
34	조상우	12000	-1302.491649
35	김재윤	12000	13651.868873
36	김세현	12000	82667.738943
37	김대현	12000	-4538.661981
38	신정락	12000	-9394.560583
39	배제성	11000	11198.160668
40	장민재	11000	-11442.874153
41	박준표	11000	56437.116626
42	박세웅	11000	17296.453184
43	박상원	11000	-5481.466464
44	장시환	11000	-4305.424947
45	배재환	10500	-4616.613043
46	전유수	10000	1290.582451
47	홍성민	10000	8802.288595
48	김원중	10000	-1699.086654
49	최성영	9500	3488.710704
50	박진형	9000	-28899.973115
51	임준섭	8800	-28955.448545
52	홍건희	8000	56611.767679
53	박치국	8000	-16139.572700
54	구승민	8000	1785.157453
55	박시영	7900	12463.312157
56	전상현	7600	44758.755411
57	이우찬	7500	-59654.345126
58	문광은	7500	-24084.268611
59	김민	7500	-1507.100945

[그림 Ⅳ-13] 투수의 실제 연봉과 예측 연봉표

V

결론 및 제언

1. 결론
2. 제언

V 결론 및 제언

1. 결론

많은 사람은 야구 경기를 볼 때 누구나 자기가 응원하는 팀이 이기기를 원하고, 응원하는 선수가 마운드에서 좋은 모습을 보여주기를 원한다. 기대에 부응하기 위해 선수들과 팀 내부에서는 팀의 승리를 위하여 다방면으로 노력하고 있으며, 선수 개개인의 연봉은 노력이 빚어낸 경기의 결과에 따라 결정된다. 연봉 인상을 마다하는 선수는 없을 것이고, 팀 내에서도 다양한 데이터 분석을 통해 객관적인 선수들의 연봉을 제안하기 위해 전력을 기울일 것이다.

본 프로젝트에서는 각 선수의 역량을 수치화한 3개년의 세이버 스탯과 이전 2개년의 연봉 데이터를 이용하여, 2020년의 연봉을 종속변수로 이를 예측하는 회귀분석 모델을 제안해보았다. 투수와 타수의 세이버 스탯이 다르므로, 선수를 투수와 타수로 이원화하여 데이터 전처리를 통해 필요 데이터를 정제하였다.

타자에서는 WAR(선수 대비 승리 기여도), wRC(wOBA 기반 득점 생산), wRAA(리그 평균 대비 득점), wOBA(출루율 스케일), SALARY(2018)(2018년 연봉), SALARY(2019)(2019년 연봉)을 독립 변수로, 투수는 WAR(선수 대비 승리 기여도), FIP(수비 무관 추정 평균 자책점), LOB%(잔루율), BABIP(인플레이된 타구의 타율), SALARY(2018년 연봉), SALARY(2019)(2019년 연봉)을 독립 변수로 사용하였다. 또한 모든 독립 변수와 종속변수의 값이 연속적이기 때문에 머신 러닝의 방법론 중 회귀분석을 사용하여 예측 모델을 구축하였다. 수정 결정 계수(R2 score)를 측정하는 방법을 이용하여 모델을 평가한 결과, 타자와 투수 각각의 회귀분석 모델을 검증하였다.

예측 모델을 통해 도출된 2020년 예측 연봉과 실제 연봉을 비교해 본 결과 다음과 같은 결론을 도출할 수 있다.

첫째, 투수 연봉 예측 모델이 타자 연봉 예측 모델보다 정확도가 높다는 것을 알 수 있다. 투수 연봉 예측 모델의 데이터의 수가 타자 연봉 예측 모델의 데이터 수보다 적음에도 이러한 결과가 나타나는 이유는 독립 변수 중 가장 유의미한 변수인

WAR의 P > |t| 값에서 찾을 수 있다. 투수 연봉 모델에서의 WAR의 P > |t| 값은 0.00인 반면, 타자 연봉 모델에서의 WAR의 P > |t| 값은 0.023이다. P > |t| 값이 낮을수록 독립 변수가 영향력 있는 변수로 볼 수 있다.

둘째, 연봉 예측 모델의 값이 음수, 즉 이상치로 나타나는 경우가 있다. 이는 회귀분석을 기반으로 한 연봉 예측 모델이 선수 개인의 세이버 스탯과 이전 2개년의 연봉이라는 객관적 데이터만을 독립 변수로 분석했기 때문이다.

현실에서 일어나는 연봉 협상은 선수의 역량 이외에도 전관예우, 선수에 대한 기대감 등의 주관적인 평가가 포함되기 때문에, 급격한 연봉 조정, 특히 연봉 삭감은 발생하기 어려운 경우가 많다. 그러나 현실적인 제약에서도 모델을 통해 도출한 예측 연봉은 연봉 협상을 위한 기준점으로 활용할 수 있다.

셋째, 타자와 투수 모두 WAR과 연봉과의 상관관계가 가장 명확하다는 것을 알 수 있다. 즉, 타자의 WAR 변수의 P > |t| 값이 0.023, 투수의 WAR 변수의 P > |t| 값이 0.00으로 모든 변수의 수치 중 가장 낮기 때문이다. 따라서 WAR 변수는 타자와 투수의 세이버 스탯 중 연봉에 가장 영향력 있는 스탯이라고 간주할 수 있으며, 연봉 상승을 위해서 선수는 WAR의 값을 높여야 한다.

2. 제언

야구는 타 스포츠보다 수치화하기 쉬운 평가 지표를 가지며, 분석 방법이 다양한 스포츠이다. 본 프로젝트에서는 세이버 스탯으로 평가한 선수 개인의 역량과 이전 연봉이 향후 연봉에 어떤 영향을 미치는가를 분석하였다.

개인의 역량과 연봉의 상관관계를 분석하고, 예측 모델을 제시하여 논리적으로 연봉을 제시할 수 있다는 함의를 가진다. 실제로 연봉 협상에서 합의점을 찾지 못해 한국야구위원회(KBO)에 연봉 조정 신청을 한 사례가 발생한 적이 있다. 이러한 상황에서 구단 측 혹은 한국야구위원회에서 연봉 예측 모델을 사용하여 객관적인 시각 자료와 지표를 제시한다면, 연봉 협상 과정에서의 시간 및 감정 소모를 단축할 수 있다. 또한, 행정적・절차적 비용의 낭비를 줄이고 선수와 구단의 편의를 도모할 수 있다.

또한, 한국 프로 야구에 관한 관심과 더불어 야구 관련 모바일과 PC 게임이 성행하고 있다. 실제로 ‘아웃 오브 더 파크 베이스볼(이하 OOTP 베이스볼)’, ‘컴투스 프로야구 2021’, ‘프로야구 매니저’, ‘마구마구’ 등 다수 회사에서 관련 게임을 서비스 중이다. 현재 구축한 연봉 예측 모델 알고리즘을 통해 게임을 개발할 때 필요한 선수의 개별 능력치, 등급, 가치, 연봉 등의 정보를 도출한다면 시간과 비용을 단축할 수 있다.

한편, 본 프로젝트는 야구 중 타자와 투수에 대한 데이터만을 활용하여 분석을 시행하였다. 그러나 이러한 변수 이외에도 포수나 경기장, 당일의 날씨 등 추가 변수가 경기와 선수의 능력치에 영향을 미칠 수 있다. 또한, 현재 사용한 변수 중 WAR을 제외한 세이버 스탯 변수가 연봉을 결정하는 완전한 변수로 간주하기에는 한계가 있기 때문에 변수를 추가한 모델을 후속 프로젝트로 발전할 수 있다. 이러한 후속 프로젝트를 통해 선수가 연봉 상승을 위해 어떤 지표를 눈여겨봐야 할 것인지에 대한 지표가 될 수 있을 것이다.

참고 논문

- 이장택, 조현식. (2009). 한국프로야구에서 데이터마이닝을 이용한 팀대 팀 승패 모형.11(6), 3417-3426.
- 오윤학, 김한, 윤재섭, 이종석. (2014). 데이터마이닝을 활용한 한국프로야구승패예측모형 수립에 관한 프로젝트. 대한산업공학회지, 40(1), 8-17.
- 장진희, 문춘걸 (2014). 한국 프로야구의 구단 승률에 대한 분석. 한국스포츠산업경영학회지, 19(3), 17-31
- 이장택. (2016). 한국프로야구에서 승률 추정방법들의 비교. 한국데이터정보과회지, 27(6), 1585-1592.
- Kim, S.-K., & Lee, Y.-H. (2016). The estimation of winning rate in Korean professional baseball league. Journal of the Korean Data and Information Science Society, 27(3), 653-661.
- 노언석. (2017).“인공신경망을 이용한 KBO 프로야구 승부예측 프로젝트.“ 국내석사학위논문 숭실대학교 소프트웨어특성화대학원, 서울

집필진

김진수	데이터 수집 및 분석
이예리	기획, 데이터 수집
임재균	데이터 전처리 및 분석
최정원	기획, 데이터 분석
홍기현	데이터 수집 및 전처리

※ 본 프로젝트 결과는 글로벌아이티인재개발원에서 빅데이터 분석가 과정을 참여하며 미니 프로젝트로 시행한 [스포츠 분석] 투수와 타자를 중심으로 본 야구 데이터 분석 및 머신러닝 의 최종보고서임.