선형 회귀 및 랜덤 포레스트를 이용한 개인 기록 기반 프로야구 선수 연봉 예측

이경문○ 황규백
숭실대학교 컴퓨터학부
leekm8474@gmail.com, kbhwang@ssu.ac.kr

Predicting salary of professional baseball players based on seasonal records using linear regression and random forest

Kyung-Moon Lee○ Kyu-Baek Hwang
School of Computer Science and Engineering, Soongsil University

요 약

1982년 KBO가 창설되고부터 야구 산업은 꾸준히 성장하여 지난 2016년에는 선수의 연봉도 구단별 상위 27명의 평균 연봉이 2억을 돌파하였으며 신인과 외국인을 제외한 기존 선수들의 평균 연봉 역시 전년도 대비 12.5% 상승하는 등의 많은 투자가 이루어지고 있다. 이에 대한 역효과로 특정 프로야구 선수들의 몸값이 지나치게 높게 측정되면서 선수 간의 연봉 편차가 커지며 논란도 점점 더 많아지고 있다. 본 논문은 연봉 예측에 선형 회귀(Linear Regression)와 랜덤 포레스트(Random Forest)를 적용하여 기록에 기반한 선수의 연봉을 예측해보았다. 그 결과로 두 기법 모두 최대 5달러 미만의 연봉오차를 보였고 랜덤 포레스트가 선형 회귀보다 조금 더 좋은 결과를 내었지만 유의미한 차이로 보기는 어려웠다. 또한 기존 유사 연구에 비해 뛰어난 성능을 보임을 실험을 통해 확인하였고 그 결과는 약 18배 더 낮은 오류율을 나타내었다.

1. 서 론

프로야구 산업의 규모가 커지면서 동시에 매년 고액의 연봉을 받는 선수들에 대한 '거품 논란' 역시 많아지고 있다. 야구단의 수익은 한정되어있는데 지나치게 높은 연봉을 받는 선수가 생기면 주어진 예산 안에서 다른 선수들의 연봉은 필연적으로 줄어들며 선수들 간의 연봉편차를 높이면서 스포츠 산업에 악영향을 주게 된다. 특히 국내의 경우, 구단의 수익이 73%나 모기업에 의존하고 있으므로 이에 대한 문제가 더 클 수밖에 없다.

본 논문은 선형 회귀와 랜덤 포레스트를 통해 선수의 시즌 기록 데이터를 분석하여 연봉을 예측하는 모델을 제시한다. 데이터 내의 어떤 지표가 연봉을 예측하는데 중요하게 반영되었는지 각각의 기법에 대해서 다뤄보고 또한 각 기법간의 성능도 비교해보았다. 마지막으로 기존에 선수의 연봉을 예측하는 연구[1]와의 비교를 진행한다.

2. 관련 연구

프로야구 선수의 연봉을 예측하는 기존의 연구로는 해당 연도의 시즌 기록을 나타내는 연간 통계정보와 해당 선수의 통산 시즌 기록을 나타내는 경력 통계정보 각각을 기반으로 진행한 연구가 있었다. 결과로, 실험 데이터의 약 96%에 대해서 오차율 0~12%의 결과를 보였다[1].

3. 시즌 기록 데이터를 활용한 연봉 예측 모델

3.1 데이터

Selenium[2]은 웹 브라우저 자동화 라이브러리로서, 본 논문에서는 Python으로 구현된 Selenium을 통해 메이저리그 시즌기록 데이터와 연봉 데이터를 수집하였다.

변수	내용
G	출장 경기
R	득점
H	안타
2B	2루타
3B	3루타
HR	홈런
RBI	타점
BB	볼넷
SO	삼진아웃
SB	도루
CS	도루실패
AVG	타율
OPS	출루율+장타율
IBB	고의사구
HBP	사구
SAC	희생번트
SF	희생플라이
TB	총 루타
XBH	장타
GDP	병살타
PA	타석
CONTRACT	계약기간
log_SALARY	당해 연봉의 자연로그 변환 값

표 1. 데이터를 구성하는 변수

위의 표 1은 수집하여 분석에 사용할 정보를 나타낸다. 본 연구는 기계학습에 사용할 데이터로 1987년~2016년까지 30년에 걸친 메이저리그 연봉 데이터[3]와 시즌 기록 데이터[4]를 선정하였고 그 중에서도 규정 타석을 채운 선수들의 데이터가 총 4531개로 수집되었다.

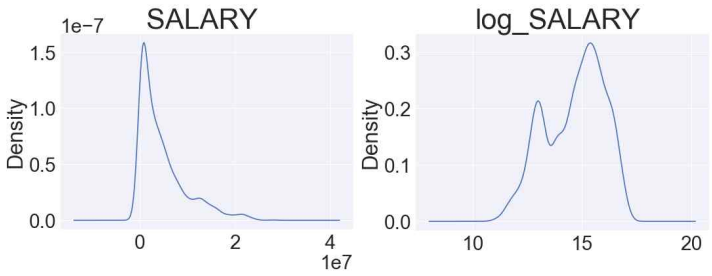

그림 1. 연봉 및 연봉의 자연로그 데이터에 대한 밀도그림

이때 분산이 커 정규분포와 매우 다른 경향을 보인 연봉 데이터의 경우, 위의 그림 1과 같이 자연로그를 통한 데이터 변환을 거쳐서 반응변수로 활용하였다.
본 논문에서는 실험을 위해 Python 기반의 'Scipy'[5]와 'Sklearn'[6] 기계학습 패키지를 사용하였다.

3.2 Linear Regression을 통한 연봉 예측 모델

아래의 표 2는 만들어진 선형 회귀 모델에서 p-value가 0.05 미만으로, 계수 값이 통계적으로 유의미하다고 볼 수 있는 설명변수 19개를 계수 절대 값의 크기순으로 나타내었다.

	Coefficient	p-value
AVG	30.3561	0.0
OPS	5.3485	0.003
CONTRACT	0.2311	0.0
SAC	-0.0603	0.0
H	-0.0515	0.0
CS	-0.0513	0.0
HR	0.0337	0.0
G	-0.0329	0.0
PA	0.0316	0.0
3B	-0.0279	0.0
BB	-0.0253	0.0
SF	-0.0249	0.002
HBP	-0.0224	O0
GDP	0.0177	0.0
TB	-0.0138	0.0
IBB	0.0125	0.0
2B	-0.0076	0.002
SB	0.0065	0.003
SO	-0.0022	0.005

표 2. 선형 회귀 모델 설명변수들의 계수

연봉에 가장 많은 영향을 주는 설명변수로는 'AVG', 'OPS', 'CONTRACT'가 각각 30.3561, 5.3485, 0.2311의 값을 가진다. 이는 각각 타율, 출루율+장타율 그리고 계약기간이다. 'AVG'와 'OPS'는 경기 내적인 기록으로 높을수록 가치가 높은 선수임을 직관적으로 이해할 수 있고 'CONTRACT'는 경기 외적인 데이터로서, 보통 고액의 연봉을 받는 선수는 구단에서 장기계약을 체결하는 경향이 있으므로 상대적으로 높은 양의 계수 값을 가지는 것으로 볼 수 있다.
반대로 직관적인 예상과는 반대의 결과가 나온 설명변수로는 '2B', '3B', 'BB'와 'TB'가 뽑혔는데 이는 2루타, 3루타, 볼넷 그리고 총 루타 값으로 양의 상관관계가 예상되었지만 모델에서의 계수 값은 모두 음의 값이 나옴을 확인되었다.

3.3 Random Forest를 통한 연봉 예측 모델

선형 회귀는 선형적 모델로서, 데이터 간의 관계 경향성을 보는데 좋지만 예측의 정확도에 대한 한계가 있다. 따라서 본 논문에서는 랜덤 포레스트 기법을 통해서도 연봉을 예측해보았다.

랜덤 포레스트는 분할하는 규칙에 따라 분할되면서 트리로 요약되고 해당 트리들을 통해 관측 값에 대한 예측 값을 연결해주는 의사결정트리(decision tree) 기법 중 하나로, 훈련 데이터 집합에서 데이터를 복원 추출하는 부트스트랩 방법을 통해 다수의 의사결정트리를 만들어내고 각 트리 내에서 분할이 고려될 때마다 모든 설명변수(p개)가 아닌 그 보다 작은 개수(m개 <p)의 설명변수(일반적으로 전체 설명변수 개수의 제곱근)로 구성된 랜덤 표본이 분할 후보로 선택되고 그 중 하나의 설명변수로 분할을 진행하여 최종 예측 값을 내놓는다[7].

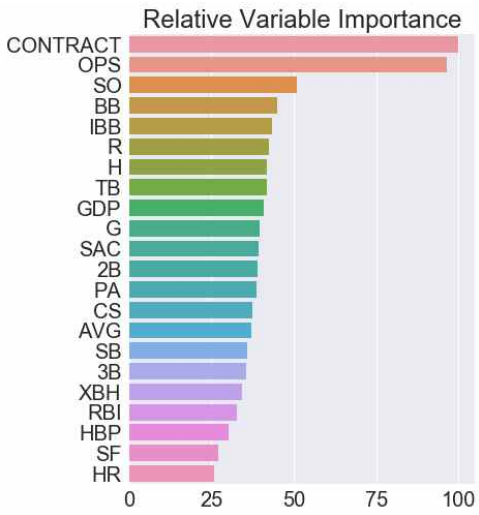

그림 2. Random Forest 모델에서 변수의 상대적 중요도

위의 그림 2는 랜덤 포레스트를 적용하면서 트리 분할에 참여한 비율에 따라 변수의 중요도를 나타낸 그래프이다. 중요도가 높은 변수로는 'CONTRACT', 'OPS'가 뽑혔는데 이 두 변수는 선형 회귀에서도 계수의 절대 값이 높은 변수로 뽑혔다.

3.4. 선형회귀와 랜덤 포레스트의 비교

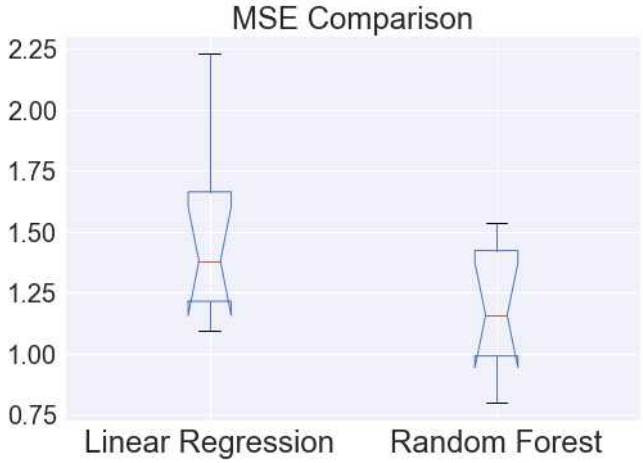

그림 1. 모델별 MSE에 대한 Box Plot

먼저 위의 그림 3은 선형 회귀와 랜덤 포레스트를 10-fold Cross Validation한 MSE를 box plot으로 나타낸 결과이다. MSE의 최대값과 중앙값은 선형 회귀에서 2.22, 1.37, 랜덤 포레스트에서 1.53, 1.15를 나타내었다. 이는 연봉으로 환산하면 선형 회귀와 랜덤 포레스트 각각 평균적으로 최대 4.43달러, 3.44 달러의 오차 내에서 연봉을 예측한다고 말할 수 있다.

랜덤 포레스트가 선형 회귀보다 조금 더 좋은 성능을 나타냈지만 그 차이가 작고 중앙값의 신뢰구간이 겹치므로 유의미한 차이라고 보기는 어렵다.

3.5 기존 유사 연구와의 비교

	기존 유사 연구	선형 회귀	랜덤 포레스트
0~3%	53.8%	33.3%	100%
3~6%	31.6%	16.6%	0%
6~9%	9.4%	28.5%	0%
9~12%	1.6%	14.2%	0%
12~15%	1.6%	4.7%	0%
15%	1.6%	2.3%	0%

표 3. 기존의 연구 모델과 본 연구 모델의 Percent-Error(%) 분포 비교

$$Percent-Error = (|실제값 - 예측값|/실제값)*100$$

수식 1. Percent-Error

기존 유사 연구에서는 2013년 타자 데이터 중 무작위 180명을 데이터로 하여 선수 개인의 통산 기록을 나타내는 경력 통계정보를 설명변수로 취하였고 본 논문에서는 2013년 규정 타석을 채운 모든 타자 선수들의 데이터로 비교하였으므로 직접적인 비교는 어려울 수 있다.

Percent-Error는 수식 1로 나타낼 수 있고 이는 예측값의 오차율로 생각할 수 있다. 위의 표 3는 기존 유사 연구 결과와 본 논문의 선형회귀/랜덤 포레스트 결과에 대한 Percent-Error 분포를 비교하고 있다. 평균 Percent-Error는 본 논문의 선형회귀가 5.94%, 랜덤 포레스트가 0.19%를 나타냈다. 또한 기존 유사 연구 Percent-Error의 평균 중앙값은 3.59%가 나오므로 랜덤 포레스트를 통해 약 18.89배 더 낮은 오차율을 보임을 간접적으로 비교할 수 있다.

4. 결론

본 연구는 두 가지 기계학습 기법들을 이용해 메이저 리그 시즌 기록 데이터와 연봉 데이터를 분석하여 중요한 변수로 계약기간과 OPS가 중복하게 뽑혔음을 알아보며 연봉 예측 모델을 제시하였다. 또한 랜덤 포레스트를 적용했을 때, 기존 유사 연구보다 더 나은 결과를 나타냄을 간접적으로 확인하였다.

본 논문에서 제시한 모델을 통해 기록으로 선수가 지니는 가치를 평가해보는 시도를 해볼 수 있고 향후에는 타자 선수뿐 아니라 투수/구단에 대한 분석까지 진행하여 어떤 선수가 특정 팀에 비용에 대비하여 높은 가치를 줄 수 있는 선수인지에 대한 연구를 진행할 수 있을 것이다.

참고문헌

[1] Rhonda Magel, Michael Hoffman. "Predicting Salaries of Major League Baseball Players", *International Journal of Sports Science*, 2015

[2] Selenium. http://www.seleniumhq.com

[3] USA TODAY, http://www.usatoday.com/sports/mlb/salaries

[4] MLB.com:The Official Site of Major League BaseBall. http://www.mlb.com

[5] SciPy.org, http://www.scipy.org

[6] scikit-learn:machine learning in Python, http://scikit-learn.org

[7] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, pp (1, 63, 375), Ruby Paper, 2016 (in Korean)