데이터를 통해 살펴보는
한국 프로야구 선수 연봉 예측

숭실대학교 컴퓨터학부 20132402 이경문

leekm8474@gmail.com

https://github.com/LeeKyungMoon/kbo-salary-prediction-modeling-randomforest

목차

** 국문요약*

데이터를 통해 살펴보는 한국 프로야구 선수 연봉 예측

1982년 한국 프로야구가 개막하고 34년이 지난 지금 2016년에 '야구'는 국내에서 가장 사랑받는 스포츠라고 말할 수 있다. 관중 현황은 1982년 143만명에서 꾸준히 증가하여 2016년의 관중 현황은 833만명으로 800만 관중시대를 열었고[1] 선수들의 연봉에 있어서 구단별 상위 27명의 평균 연봉이 올해 2억원을 처음으로 돌파하였고 신인과 외국인을 제외한 기존 선수들의 평균 연봉은 1억 2656만원으로 전년 대비 12.5% 상승하는 것을 통해 알 수 있듯이 많은 투자가 이루어지고 있다.[2] 야구는 기록의 스포츠로서, KBO(Korea Baseball Organization)의 누적 데이터가 쌓여가고 또 한국 프로야구 역시 규모가 커지고 선진화되어가며 각 구단들은 데이터를 적극적으로 활용해 전략을 세우고 상대방에게 대응하고 있다.

모든 산업의 핵심 역량으로 대두되고 있는 빅데이터 분석, 머신러닝을 활용하여 한국 프로야구선수들의 기록 데이터를 분석하여 주어진 선수의 기록에 대한 선수의 연봉을 예측하는 모델링을 제시한다. 예측 모델링에는 머신러닝 분야에서도 지도학습 (Supervised Learning) 분야를 통해 진행되고 특별히 랜덤 포레스트 (Random Forest) 기법을 활용한다. 해당 분석은 데이터 분석에 많이 쓰이고 있는 프로그래밍 언어 Python 기반의 Project Jupyter'에서 진행되었고 'Pandas', 'Numpy', 'Seaborn', 'Sklearn', 'Selenium'과 'Matplotlib'패키지들을 함께 사용하였다.

---

1) 관중현황-KBO 정규시즌: http://www.koreabaseball.com/Record/Crowd/Graph_Year.aspx

2) KBO리그, 1군 평균 연봉 2억 돌파...억대 연봉 148명, <JTBC 일간스포츠>, 2016-02-11 08:53, http://news.jtbc.joins.com/article/article.aspx?news_id=NB11170960

본 프로젝트는 KBO 프로야구 선수들의 기록 데이터를 기반으로 한국 프로야구 선수의 연봉을 예측하는 모델링을 제안하여 거시적인 데이터를 통해 추상적인 진실을 탐구해보는 시도를 해보는 의의와 날로 높아져가는 KBO 한국 프로야구선수들의 연봉에 대해, 데이터를 통한 조금 더 객관적인 기준으로 한국 프로야구 선수의 연봉을 평가해보는 시도를 하는데 의의를 가진다.

1) 서론

1. 배경 및 목적

2016년 올해 800만 관중시대를 연 KBO는 2015년 기준으로 170만 관중을 기록한 프로축구나 110만 관중을 기록한 프로농구[3]에 비해서 압도적으로 많은 관중수를 확보하고 있다. 이는 프로축구나 프로농구와는 달리 이전부터 올해까지 지속적으로 꾸준히 증가한 경향을 보이고 있으므로 내년은 더 많은 관중수를 기대할 수 있다는 점에서 '야구'라는 스포츠가 대한민국에서 가장 사랑받는 스포츠라고 말할 수 있는 지표로서, 더 확고히 자리매김하고 있다고 말할 수 있다.

또한 2016년 올해는 메이저리그에서 경기한 한국인 선수가 가장 많았던 해로, 총 6명의 한국인 야구선수들이 활약하며 '저비용 고효율'의 성과를 내며 인정받기도 했다. 이렇게 한국 프로야구 선수들의 수준 역시 크게 성장하였고 그에 따라 KBO에 대한 더 많은 투자와 관심을 만들어내고 있다. 자연스럽게 이는 선수들의 연봉에도 반영되어 평균 연봉의 상승과 고액 연봉을 받는 선수들이 증가하는 결과를 만들어 냈다. 이는 건강한 순환처럼 보이지만 매해 FA 계약을 통해 초고액 연봉 계약이 성사될 때마다 전문가들과 팬들의 경계도 이어지고 있다[4].

야구는 데이터 스포츠라고 불린다. 구단 별로 존재하는 전력분석팀이 경기 내에서 발생하는 수많은 데이터를 분석하여 전략을 구상하는 것을 넘어 구단의 마케팅과 선수들과의 계약할 때의 지표로도 활용되는 등 스포츠 분야의 많은 영역에서 활용되고 있다. 특별히 한국에서는 선수의 스타성보다는 데이터와 통계를 적극 활용한 경제적인 선수 영입을 통해 스몰 마켓의 하위 구단을 성공적으로 성장시키는 내용의 영화 '머니볼'을 통해 '세이버메트릭스'라는 야구 이론과 함께 스포츠에서 데이터의 중요성이 많이 알려지기도 했다.

이렇게 다양한 분야에서 수많은 데이터가 생겨나고 IT 기술이 발달하며, 빅데이터를 저장하고 분석할 수 있게 되었다. 이에 대한 분석 기법이자 인공지능의 한 분야인 머신러닝은 컴퓨터가 스스로 복잡하고 다양한 자료에서 상관관계를 분석하고 학습하여 유의미한 결과를 내는 분야이다.

따라서 야구에서 선수가 가지는 여러 가지 데이터 항목들을 머신러닝 기법을 통해 분석하여 데이터에 따른 선수들의 연봉을 예측하는 모델을 생성하고 이에 대한 성능을 평가하고 선수의 기록 데이터에 대하여 모델이 제시하는 연봉과 실제 연봉이 어떤 차이를 보이는지 살펴보고자한다.

---

3) 국가지표체계 : http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=1662

4) [빅콘] 'FA 100억 시대' 팬들 반응은? 부정 55%>긍정 45%, <한국스포츠경제>, 2016.12.08 09:08, http://www.sporbiz.co.kr/news/articleView.html?idxno=56988

2. 범위 및 구성

본 프로젝트는 KBO 공식 홈페이지 기록실 사이트를 프로그래밍 언어 Python 기반의 'Selenium' 패키지를 통해 크롤링한 데이터를 머신러닝(Machine Learning) 기법을 'Pandas', 'Numpy' 패키지를 통해 사용하여, 선수별 기록 데이터와 선수별 연봉 관계를 파악하고 그에 따른 모델링과 해당 모델링의 성능과 한계에 대해서 알아보도록 한다.

수집한 데이터는 타자(Hitter) 선수들에 대한 데이터만 수집하였고 항목으로는 AVG (타율), G (경기 수), PA (타석 수), AB (타수), R (득점 수), H (안타 개수), 2B (2루타 개수), 3B (3루타 개수), HR (홈런개수), TB (누적루타 수), RBI (타점 수), SAC (희생번트 수), SF (희생플라이 수)와 SALARY (연봉액수, 단위:만원)으로 14개의 항목의 데이터가 수집되었다.

수집한 데이터에 대한 분석으로는 SALARY와 SALARY를 제외한 13개의 항목에 대한 상관관계를 그래프를 통해 먼저 살펴본다. 이후에 다중회귀분석에 대해 기술한 뒤, 해당 분석을 진행하여 각 항목이 SALARY에 어느 정도 영향을 끼치는지 살펴보고 또 다른 분석 방법인 트리 기반(Tree-based)의 방법과 그것의 일종인 랜덤 포레스트(Random Forest) 기법에 대해 기술하고 이에 대한 분석을 진행하여 생성한 예측 모델링에 대한 성능을 평가해본다. 마지막으로 해당 예측 모델링에 대한 의의, 한계 그리고 개선해야할 부분에 대해서 기술하며 마무리한다.

2) 본론

1. 통계학습

통계학습은 데이터에 대한 이해를 위한 방대한 도구 집합을 말하고 이에 대한 종류로는 지도학습과 비지도학습으로 분류될 수 있다.[5] 본 프로젝트는 데이터 분석을 통해 한국 프로야구 선수의 연봉이라는 수치를 결과로 내놓는 것으로 하나 이상의 입력변수를 기반으로 출력변수를 예측 및 추정하는 지도적 통계학습에 해당된다.

2. 다중선형회귀분석

회귀분석이란 통계학적으로, Y로 표시하고 반응변수라고 불리는 값과 X로 표시하고 설명변수라고 불리는 값 사이의 관계를 모델링하는 기법이다[6]. 이때 단순회귀분석 ($Y = aX + \alpha$)은 설명변수 X가 하나인 경우이고 다중회귀분석 ($Y = aX1 + bX2 + \cdots + nXn + \alpha$)이란 설명변수가 2개 이상 쓰이는 것으로 설명변수 X 앞에 붙은 계수(a,b,…,n)는 해당 설명변수의 단위가 하나 증가할 때마다 평균적으로 증가하는 반응변수 Y의 평균 증가량으로 해석할 수 있다.

---

5) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R (가볍게 시작하는 통계학습 : R로 실습하는), 마이클, 루비페이퍼, 1쪽

6) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R (가볍게 시작하는 통계학습 : R로 실습하는), 마이클, 루비페이퍼, 63쪽

그렇다면 데이터를 통해 만든 모델의 계수(a,b,...,n)는 어떻게 추정해야하는지에 대한 기준이 필요한데 그것은 잔차제곱합(Residual Sum of Squares)이다. 본 프로젝트에 대응하여 잔차제곱합을 설명하자면, 동일한 설명변수 조건 하에서, 수집한 데이터 집합 중 I번째 선수의 연봉(pi)과 해당 모델이 제시하는 연봉의 추정량(qi)의 차의 제곱의 합이다.

따라서 데이터의 개수가 n개일 때 이를 수식으로 나타내면 $\sum_{i=1}^{n} (\pi - qi)^2$ 로 나타낼 수 있다.

그리고 주어진 데이터 집합은 하나로 정해져있지만 해당 데이터 집합에 대한 모델은 여러 가지가 나올 수 있다. 이때 최소의 잔차제곱합을 가지는 모델을 최적의 모델로 선택하는 것이다[7]. 따라서 본 프로젝트는 연봉과 각 항목 간의 도표를 그린 뒤 도표를 통한 해석과 직관적인 판단을 비교해보고 한국 프로야구 선수의 연봉을 반응변수 Y로 두고 나머지 카테고리 데이터를 설명변수 X로 둔 뒤 다중회귀분석을 진행해 보고자한다.

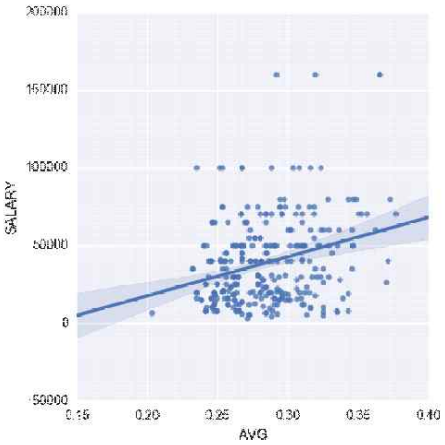

그림 1 : 연봉과 타율의 상관관계

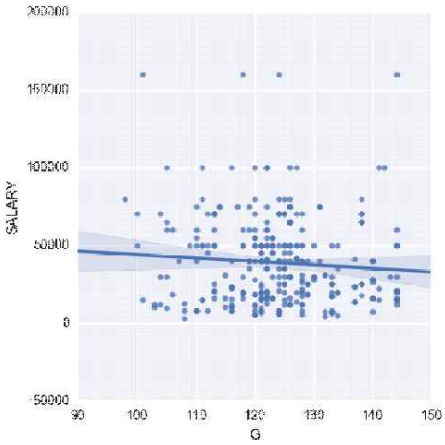

그림 2 : 연봉과 게임 수의 상관관계

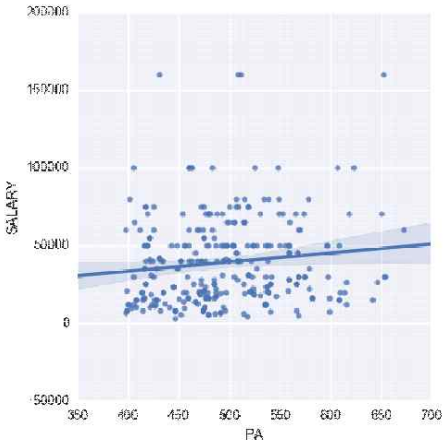

그림 3 : 연봉과 타석의 상관관계

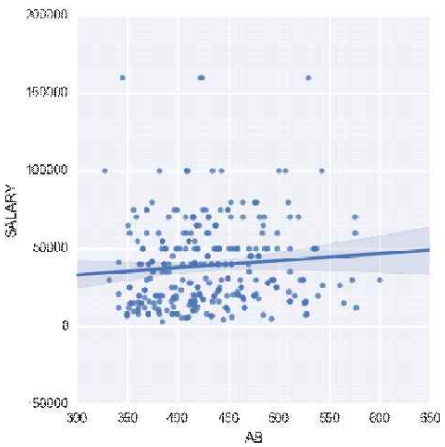

그림 4 : 연봉과 타수의 상관관계

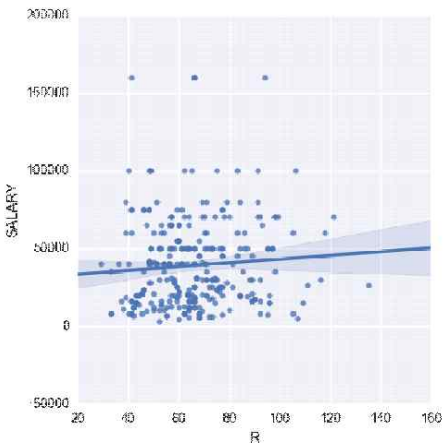

그림 5 : 연봉과 득점의 상관관계

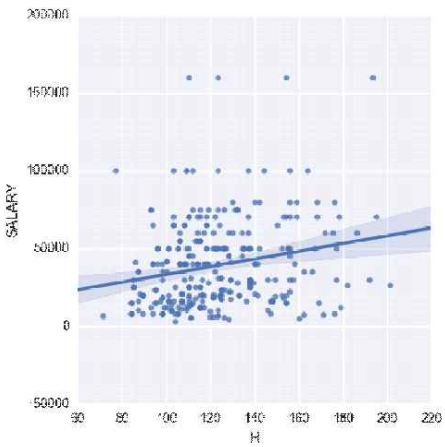

그림 6 :연봉과 안타 개수의 상관관계

7) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R (가볍게 시작하는 통계학습 : R로 실습하는), 마이클, 루비페이퍼, 66쪽

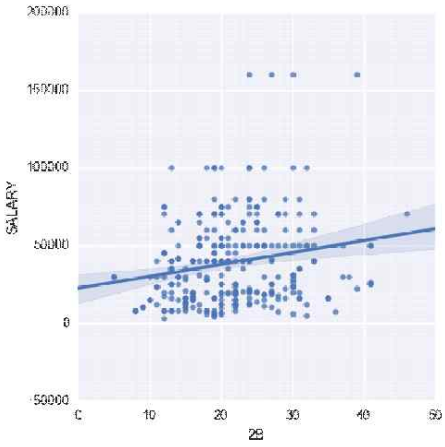

그림 7 : 연봉과 2루타의 상관관계

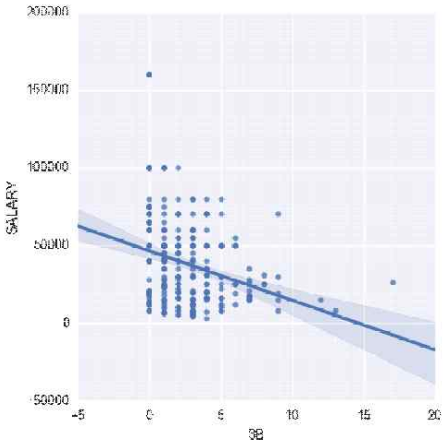

그림 8 : 연봉과 3루타의 상관관계

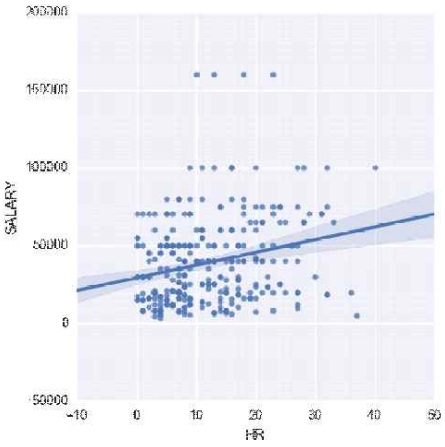

그림 9 : 연봉과 홈런의 상관관계

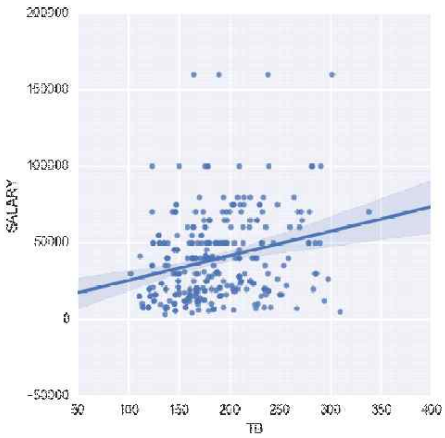

그림10 :연봉과 누적루타의 상관관계

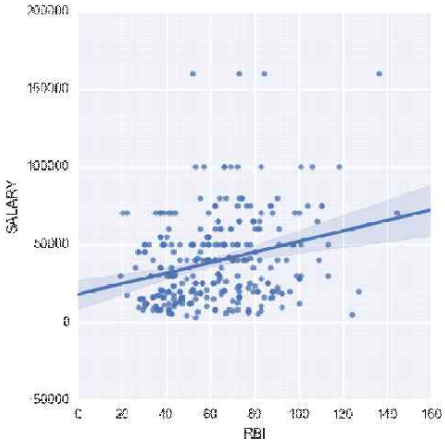

그림 11 : 연봉과 타점의 상관관계

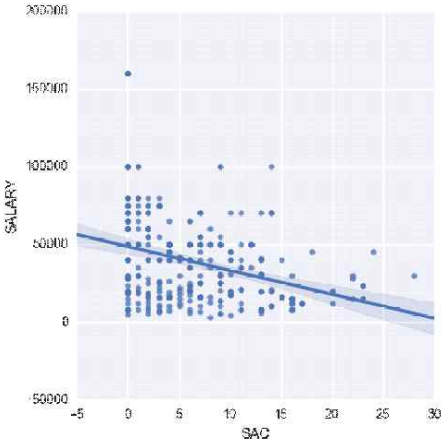

그림12 :연봉과 희생번트의 상관관계

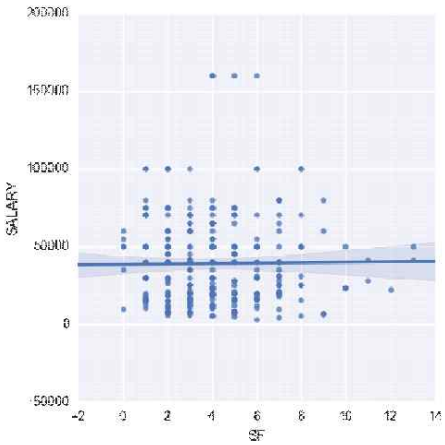

그림13:연봉과 희생플라이의 상관관계

13개의 차트 모두 Y축은 선수들의 연봉인 SALARY를 가리키고 X축은 각각 {타율, 게임수, 타석, 타수, 득점, 안타, 2B, 3B, HR, 누적루타, 타점, 희생번트, 희생플라이}를 가리키면서 두 값의 상관관계를 도표로 나타내고 있다. 도표를 통해 {타율, 안타, 2루타, 홈런, 누적루타, 타점}의 도표는 다른 도표에 비해 높은 양의 상관관계를 가짐을 확인할 수 있고 이는 도표를 그리지 않아도 직관적으로 이해할 수 있는 부분이다. 그 밖의 지켜볼 부분으로는 {'3루타', '희생번트'}는 연봉과 상대적으로 높은 음의 상관관계를 가지고 있음을 확인할 수 있는데 3루타는 2루타보다 좋은 기록임에도 데이터 상으로는 연봉과 음의 상관관계를 가진다는 부분이 직관적인 판단과는 다르게 나타는 것을 확인할 수 있다. '연봉'에 영향이 거의 없는 항목으로는 '희생플라이'가 도표 상으로 그렇게 보임을 알 수 있다. 마지막으로 {타석 수,

타수와 득점}은 연봉과 상대적으로 작은 양의 상관관계를 보이고 '게임 수'는 연봉과 상대적으로 작은 음의 상관관계를 보여주고 있다. 기울기의 절대값이 상대적으로 작음지만 게임 수가 많아지면 타석과 타수도 증가하는데 '게임 수'와 {'타석', '타수'}는 서로 배치되는 상관관계를 나타내고 있다는 점도 지켜볼만 하다.

이어서 연봉을 반응변수로 두고, 나머지 지표들을 설명변수로 둔 뒤에 다중선형회귀분석을 진행한 결과로는 다음과 같은 수식이 나오게 된다.

$$\begin{aligned} ALARY \quad & 2.5\times 10\ \times(AVG)-6.95\times 10^2\times(G)+4.1\times 10^2\times(PA)-2.6\times 10^2\times(AB) \\ & -2.62\times 10^2\times(R)+3.34\times 10^2\times(H)+5.97\times 10^2\times(2B)-2.37\times 10^3\times(3B)+1.14\times 10^3\times(HR) \\ & -4.02\times 10^2\times(TB)+1.62\times 10^1\times(RBI)-7.79\times 10^2\times(SAC)-8.72\times 10^2\times(SF) \end{aligned}$$

위에서 살펴본 도표에서와 비교해볼 때, 타율이 가장 큰 양의 상관관계를 가지는 점에서 동일하지만 '희생 플라이'는 도표 상으로 볼 때는 연봉과 상관관계가 없는 것처럼 보였지만 다중선형회귀분석을 한 결과에서는 연봉과 상대적으로 큰 음의 상관관계를 나타내고 있음을 알 수 있다. 이와 같이 실제로 단순회귀계수와 다중회귀계수는 상당히 다를 수 있는데 이는 단순회귀의 경우 다른 설명변수에 대한 고려가 없고 또 여러 가지 변수들 간에 존재하는 시너지 또는 상호작용 효과에 대한 고려가 없기 때문이다[8]. 이에 대해 각 설명변수들이 적합한지를 보여주는 지표는 각 설명변수들에 대한 p값이다. p값은 사전적 의미로는 검정 통계량이 실제 관측된 값보다 대립가설을 지지하는 방향으로 더욱 치우칠 확률로, 귀무가설 하에서 계산된 값이다. 이때 귀무가설은 해당 설명변수의 값이 0으로 반응변수에 대한 영향이 없다는 내용이 되고 95% 신뢰도 기준에서 p값이 0.05보다 크면 귀무가설을 채택하면서 해당 설명변수는 반응변수에게 영향이 없다고 보게 된다.

위의 다중회귀분석에 대한 각 변수들에 대한 통계량은 오른쪽에 나와 있는 표를 통해 확인할 수 있다. 맨 오른쪽의 값이 p값으로, 변수 {AB,R,H,RBI,SF} 들은 0.05를 넘으므로 통계적으로 유의미하지 못한 변수라고 볼 수 있고 해당 변수들은 제외시킬 수 있는 변수들로 볼 수 있다. 이렇게 단순하게 다중 선형 회귀 분석을 진행하였을 때는 각 설명변수들의 분산도 높고 그에 따른 최종 성능이 안 좋기 쉽다. 따라서 이를 개선하기 위해 트리 기반의 방법 중에서도 랜덤 포레스트(random forest) 기법을 활용하여 예측 정확도를 높여보도록한다.

	coef	std err	t	P>|t|
AVG	2.489e+05	6e+04	4.148	0.000
G	-695.0198	220.923	-3.146	0.002
PA	410.0392	109.042	3.760	0.000
AB	-259.8102	133.514	-1.946	0.053
R	-261.6078	170.444	-1.535	0.126
H	334.1791	224.090	1.491	0.137
2B	596.9414	298.452	2.000	0.046
3B	-2369.5272	547.520	-4.328	0.000
HR	1135.1836	288.247	3.938	0.000
TB	-402.3831	137.693	-2.922	0.004
RBI	16.1644	163.709	0.099	0.921
SAC	-778.5485	307.857	-2.529	0.012
SF	-871.8222	677.192	-1.287	0.199

8) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R (가볍게 시작하는 통계학습 : R로 실습하는), 마이클, 루비페이퍼, 80쪽

3. 랜덤 포레스트

“트리기반방법” :

트리 기반 방법은 설명변수의 공간을 다수의 영역으로 계층화 또는 분할하는 방법이다. 분할하는 규칙에 따라 분할되면서 트리로 요약되고 해당 트리를 통해 관측값에 대한 예측값을 연결해주는 예측 모델링 방법이고 이러한 유형의 기법들을 의사결정트리(decision tree)이라 부른다(결정 트리 중 목표 변수가 연속하는 값, 일반적으로 실수를 가지는 것은 회귀 트리라 한다)[9].

위에서 진행한 다중선형회귀분석과 비슷하게 처음에 트리를 만들어 나갈 때는 재귀이진분할(recursive binary splitting)로 알려진 하향식의 그리디 기법을 통해 최대한 많은 경우를 고려하는 트리를 형성하고 그런 뒤에 결과 트리가 너무 복잡하여 오히려 성능이 떨어질 수 있는 ‘과적합’ 현상이 생기는 것을 방지하기 위해 가지치기(tree pruning)을 진행한다[10].

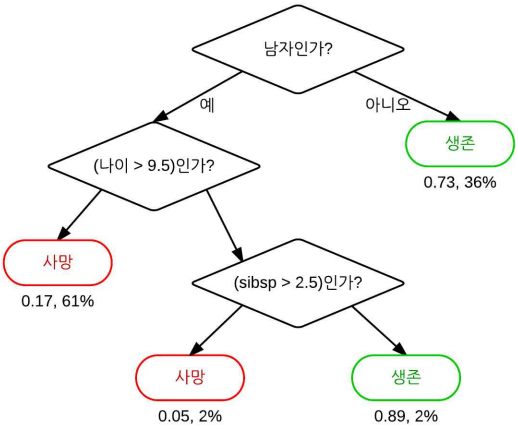

그림 14 : 결정트리학습법

이를 그림[11]으로 나타내자면 오른쪽의 그림과 같이 주어진 트리에 예측하려는 데이터 모델을 넣으면 트리를 따라 내려오면서 결과값을 예측하는 기법이다.

이때 랜덤 포레스트 방법은 해당 결정트리학습법에 대한 더 강력한 예측모델을 구성하는 기법 중 하나로서, 부트스트랩 방법을 통해 훈련 데이터 집합에서 다수의 의사결정트리를 만들어내고 각 트리 내에서 분할이 고려될 때마다 모든 설명변수(p개)가 아닌 그 보다 작은 개수(m개 < p)의 설명변수(일반적으로 전체 설명변수의 개수의 제곱근)로 구성된 랜덤 표본이 분할 후보로 선택되고 그 중 하나의 설명변수로 분할을 진행한다[12].

랜덤 포레스트 방법의 특징 중 하나가 트리가 내려가면서 분할될 때, 전체가 보다 적은 m개의 설명변수를 전체 설명변수에서 랜덤하게 뽑아서 분할의 기준으로 뽑는다는 것이다. 이것의 장점은 대부분 많은 트리에서 중요도가 높은 설명변수가 분할에서 큰 영향을 미치기 마련인데, 사실은 이것이 자연스러워 보이지만 예측치들이 서로 높은 비율로 상관하게 되면서 성능상으로는 좋지 못한 결과를 내는데 이를 랜덤하게 분할하면서 중요도가 높은 설명변수가 분할에서 공평하게 영향을 미치면서 예측치들간의 상관비율을 낮추면서 성능은 높이게 되는 결과를 낳게 되는 것이다.

---

9) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R (가볍게 시작하는 통계학습 : R로 실습하는), 마이클, 루비페이퍼, 355쪽

10) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R (가볍게 시작하는 통계학습 : R로 실습하는), 마이클, 루비페이퍼, 359~360쪽

11) 그림 14 : 결정 트리 학습법 그림, 위키백과

12) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R (가볍게 시작하는 통계학습 : R로 실습하는), 마이클, 루비페이퍼, 375쪽

본 프로젝트의 데이터 집합에 대한 결정 트리는 다음 이미지와 같다.

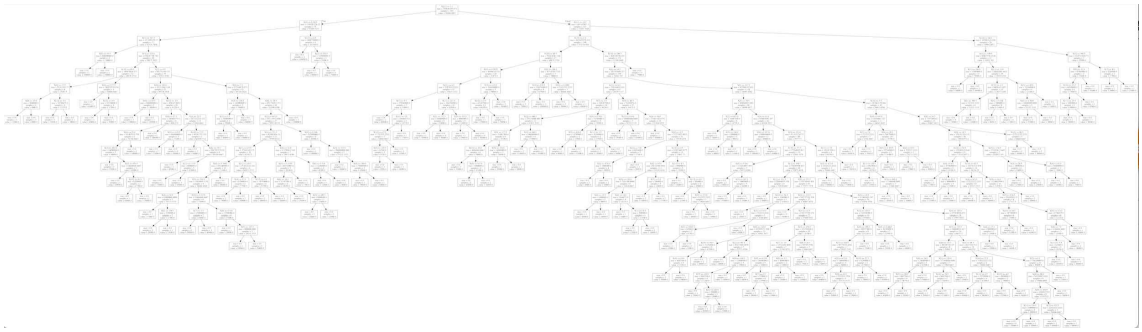

그림 15 : KBO 데이터 집합에 대한 의사결정트리 전경

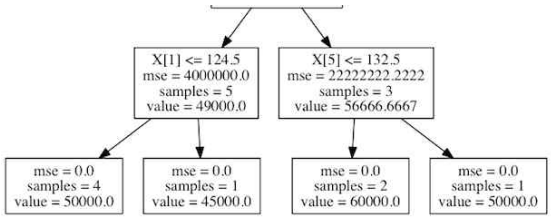

그림 16 :
KBO 데이터 집합에 대한
의사결정 트리의 말단부분

'그림 15'는 본 프로젝트에서 실시한 KBO 데이터 집합에 대한 의사결정트리 전체를 그림의 형식으로 나타낸 것이고 '그림 16'은 '그림 15'에서 오른쪽 하단의 말단 부분을 확대한 사진으로 왼쪽에서부터 연봉에 대한 추측 값을 50000, 45000, 60000, 50000(단위: 만원)으로 적혀있음을 확인할 수 있다.

다음은 KBO 데이터 집합에서의 연봉(SALARY)과 랜덤 포레스트 기법을 이용하여 추측한 연봉(PREDICTION)을 비교하는 이미지이다.

	SALARY	PREDICTION
0	70000	74000.0
1	160000	143000.0
2	70000	65500.0
3	50000	53000.0
4	80000	89000.0
5	8000	22400.0
6	19000	27000.0
7	60000	58000.0
8	50000	45700.0
9	7000	12600.0
10	7700	8560.0
11	5000	34500.0
12	10000	26000.0
13	7000	20500.0
14	26000	31400.0

그림 17 : KBO 데이터 집합에서의 SALARY와 KBO 데이터에 대해 랜덤 포레스트 기법을 활용한 예측 결과 PREDICTION

최종적으로 KBO 데이터 집합에 대한 랜덤 포레스트 기법의 오류율은 33%으로 67%의 성능을 보인다고 볼 수 있다.

```
error_rate = abs((predict_random_forest - kbo['SALARY']) / kbo['SALARY'])
error_rate.mean()
```

```
0.33141848060526125
```

그림 18 : 랜덤 포레스트 기법을 통한 예측에 대한 오류율을 프로그램을 통해 계산하는 그림으로 0.33으로 33%를 나타내고 있다.

'그림 17'의 첫 번째 줄 데이터는 2016년 타율 1위인 '최형우' 선수에 대한 데이터를 넣었을 때, 예측 연봉 값을 나타낸 값으로 올해 해당 선수의 연봉은 7억원이고 이에 대한 예측 값을 7억 4천만원으로 예측하였다. 아래의 '그림 19'는 해당 선수에 대한 기록 지표이다

	AVG	G	PA	AB	R	H	2B	3B	HR	TB	RBI	SAC	SF
0	0.376	138	618	519	99	195	46	2	31	338	144	0	7

그림 19 : '그림 17'의 첫 번째 데이터를 기록한 '최형우' 선수의 2016년 시즌 기록 데이터

3) 결론

1. 요약 및 의의

먼저 KBO 공식 홈페이지의 기록실의 데이터를 'Selenium' 패키지를 통해 수집하여 KBO 데이터 집합으로 만들었다.

이후에 설명 변수들{AVG, G, PA, AB, R, H, 2B, 3B, HR, TB, RBI, SAC, SF} 각각과 SALARY 데이터 간의 도표를 'Seaborn' 패키지를 사용해서 그리면서 설명변수 하나와 SALARY 반응변수 간의 전반적인 관계를 살펴보았고 이어서 모든 설명 변수들을 포함하여 반응변수인 SALARY를 추정해보기 위해 'sklearn' 패키지를 사용하여 다중선형회귀분석을 진행하여 설명변수와 반응변수에 대한 구체적인 수식을 이끌어내어 앞에서 진행한 도표와 비교해보았다. 이때 도표에서 간단하게 살펴본 상관관계와 다중선형회귀분석에서의 상관관계가 차이가 있을 수 있음을 살펴보았고 그 이유가 도표에서는 다른 변수들에 대한 고려가 없이 진행되었기 때문에 이에 대한 처리가 있은 후에는 상관관계가 다르게 나올 수 있고 또 다중선형회귀분석에서의 각 설명변수에 대한 p값이 95% 신뢰수준에서 0.05보다 크면 해당 설명변수는 목표값인 연봉 반응변수에 대한 영향력이 없다는 것을 통계적으로 의미하여 해당 설명변수를 제외시킬 수도 있음을 말했다.

마지막으로 기존의 다중선형회귀분석을 개선시키기 위해 트리기반의 방법을 통한 분석을 진행하였고 그 중에서도 랜덤 포레스트 기법을 사용하여 주어진 설명변수 데이터에 대한 반응변수인 연봉을 예측하는 모델링을 만들어보았고 해당 모델은 평균적으로 33%의 오류율을 나타내면서 예측 값을 제시하고 있음을 확인하였다.

매년 KBO의 스타 선수들이 FA 계약을 할 때마다 '몸값 거품' 논란이 이어지고 있다. 사실 한국 프로야구를 해외 야구리그와 비교해 볼 때, 평균 연봉에 있어서 메이저리그는 KBO의 약 40배, 일본 프로야구는 KBO의 약 3배를 기록하고 있다. 이때 단순 비교는 어려우므로 국가별 GDP를 고려해볼 때, 미국은 한국의 약 13배이고 일본은 한국의 약 3배의 GDP를 기록하고 있다. 따라서 국가의 시장 규모를 고려하더라도 메이저리그는 KBO 보다 많은 연봉을 기록하고 있다고 말할 수 있지만 일본 프로야구는 딱히 그렇다고 말하기가 쉽지 않다. 이 역시도 각 국가별 야구 리그의 수준에 대한 정확한 비교 없이 진행되는 것이므로 더 더욱 단순 비교하기가 어려운 부분이 있다. 하지만 리그별 구단의 수익 구조를 살펴볼 때, 메이저리그의 경우에 경기장 내 수익이 50%, 스폰서 20% 그리고 중계권료 20%를 이루는데 국내 구단의 경우에는 경기장 내 수익이 10%가 안되고 무려 70%에 해당하는 모기업 투자가 이루어지고 있다. 따라서 국내 구단이 가지는 수익의 대부분이 모기업에 의존적이라고 말할 수 있고 그때 초고액 연봉자가 생기면 정해진 예산 안에서 다른 선수들의 연봉이 필연적으로 줄어들 수밖에 없는 구조이다. 따라서 초고액 연봉자가 계속 배출되는 것은 생태계에 좋지 못한 영향을 줄 수 있다[13].

따라서 본 프로젝트는 객관적인 데이터를 통해 해당 기록 데이터를 가지고 있는 선수가 지니는 가치에 대해서 평가를 시도하면서 해당 선수가 현재 과대평가 받고 있는지, 과소평가 받고 있는 지에 대한 평가를 시도해본다는 점에서 의의를 가진다.

---

13) [팩트체크] 논란의 프로야구 선수 '몸값'…거품인가 아닌가, <JTBC>, 2015-12-01 23:30, http://news.jtbc.joins.com/article/article.aspx?news_id=NB11109153

2. 한계점과 개선방향

본 프로젝트가 단순히 도표를 그리는 것부터 시작해서 머신러닝 기법 중 하나인 랜덤 포레스트 방법을 통해 선수의 연봉을 예측하는 모델을 만들어보았지만 여러 가지 한계점을 가진다.

먼저 야구에는 수많은 데이터 지표들이 있지만 그 중 일부인 기본적인 지표들(AVG 타율, G 경기 수, PA 타석, AB 타수, R 득점, H 안타, 2B 2루타, 3B 3루타, HR 홈런, TB 누적루타, RBI 타점, SAC 희생번트, SF 희생플라이)에 대해서만 평가가 이루어졌고 투수 선수들을 제외한 타자 선수들에 대한 데이터만으로 만들어진 모델링 역시 커버하는 데이터 영역이 그만큼 좁을 수밖에 없다.

그에 대한 예시로 '연차'에 대한 기록이 빠져서 생기는 성능저하를 살펴볼 수 있다. '그림 17'에서 예측 값와 실제 값이 차이가 상대적으로 많이 나는 데이터의 경우가 6번째, 10번째, 11번째 그리고 14번째 줄 데이터가 그렇다. 이 데이터의 공통점으로 볼 수 있는 것은 4명의 선수 모두 연차가 많지 않지만 좋은 기록을 낸 선수들이다. 4명의 선수의 이름은 각각 구자욱, 박건우, 고종욱 그리고 김문호 선수인데 이들의 각각 데뷔년도는 2012년, 2009년, 2011년 그리고 2006년으로 평균 연차가 7.25년인데 2016년 기준 연봉 상위 4명의 선수가 김태균, 강민호, 최정 그리고 이승엽 선수로 평균 연차가 15.75년으로 차이가 2배 이상 나는 것을 확인할 수 있었다. 따라서 직관적으로도 이해할 수 있고 데이터를 통해서도 볼 수 있듯이 연차는 연봉에 영향을 크게 주는 요소임에도 해당 데이터가 분석에 포함되지 않아서 연차가 적지만 좋은 기록을 낸 선수들에 대한 연봉 예측 오류율이 높게 나타나고 있다.

또한 제외한 데이터들에 대해서 살펴보자면, 해당 선수가 어디 팀 소속인지도 연봉에 영향을 줄 수 있는데 선수 개인의 기록에 대한 연봉을 예측하는 것이므로 팀 소속에 대한 데이터를 제외하였고 KBO에서 현역을 은퇴한 선수들에 대한 연봉 데이터를 주지 않고 있고 외국인 선수의 경우 '달러' 단위로 계약하므로 해당 선수들에 대한 데이터는 제외하였다.

마지막으로 경기 내의 상황인 기록 데이터뿐만 아니라 해당 선수가 어떤 에이전트의 소속인지와 같은 외부적 상황 데이터도 존재한다. 기록 데이터는 고정되어 있지만 계약을 어떻게 하느냐에 따라 선수에 대한 가치를 과대평가 또는 과소평가 할 가능성이 존재한다. 따라서 해당 선수의 에이전트가 어떤 계약을 체결해왔는지에 대한 정보도 미래의 계약 성과에 대한 지표가 될 수 있고 총액이 같은 조건으로 계약하더라도 계약금와 연봉을 어떻게 나누었는지에 대한 정보도 연봉 자체에 대한 예측을 할 때 영향을 주게 되는데 이러한 외부적 상황에 대한 고려는 이루어지지 않았다.