빅데이터로 보는 야구 – 최종보고서

15학번 정우진

15학번 제갈장환

16학번 김호진

1. 소개

야구를 빅데이터로 분석하는 것은 많이 알려져 있다. 이는 영화 ‘머니볼’에서 등장한 세이버 매트릭스라고 불리고 있다. 많은 분석을 통해서, 사람들은 승리에 관한 지표나 한 사람이 어떤 능력치를 가지고 있는지에 대해 알아왔다. 이러한 지표 중에서 가장 유명한 것은 WAR(대체 선수 대비 승리 기여도)이다. 이는 선수가 팀 승리에 얼마나 기여했는가를 표현하고 있다. 이미 MLB에서는 오랜 시간 동안 누적된 자세한 데이터를 기반으로 이러한 상수를 상세히 분석하여 사용하고 있지만, 비교적 역사가 짧은 KBO에서는 이 상수를 자세하게 구하기 힘들다. 허나, 야구를 분석하고자 하는 사람들이 모인 ‘스탯티즈’와 ‘KBReport’에서 각자의 방식으로 계산을 하였다. KBReport에서는 타자의 수비율을 포함하지 않은 계산법을 사용하였고, 스탯티즈에서는 타자의 수비율이 어느 정도 포함된 계산법을 사용하였다. 또한, 실제로 선수의 WAR에 대한 정확성은 스탯티즈가 더 높은 편이기에 스탯티즈의 WAR 값을 사용하여 분석을 하기로 하였다. 이외의 데이터들은 KBO에서 제시하는 데이터를 사용하였다.

이 프로젝트에선 위의 데이터들을 기반으로, 팀의 승리를 이끌어 낼 수 있는 방법, 승리에 영향을 끼칠 수 있는 요인들 그리고 실제로 선수의 연봉을 예측해보기 위한 분석을 진행하였다.

2. 분석 과정 및 결과

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| WPCT | 승률 | AVG | 타율 | PA | 타석 |
| OBP | 출루율 | ERA | 평균 자책점 | AB | 투수 |
| OPS | 출루율+장타율 | FIP | 수비 제외 투수 능력 | TB | 루타 |
| SLG | 장타율 | HR | 홈런 | H | 안타 |
| R | 득점 | 2B | 2루타 | SV | 세이브 |
| RBI | 타점 | BB | 볼넷 | XBH | 장타 |

Table 1 지표들에 대한 설명

위 표는 분석하는 과정에 필요한 지표들이다. 이외에 더 필요한 설명에 대해서는 분석을 하는 과정에서 설명할 것이다.

1) 팀 승률을 어떻게 올릴 수 있는가?

|  |  |
| --- | --- |
| Figure 1 팀 타자 WAR과 승률 | Figure 2 팀 투수 WAR과 승률 |

이때, 사용한 데이터는 승률을 알 수 있는 1991년부터 2017년까지의 데이터를 기반으로 실행하였다. 위의 결과를 보았을 때, 타자 WAR, 투수 WAR은 팀 승률과 각각 0.755, 0.822의 Pearson correlation coefficient 를 가진다. 이는 두 가지 모두 강한 양적 선형관계를 가지게 된다. 그러므로 팀 승률과 팀 전체의 WAR의 합은 매우 연관성이 높다. 이는 팀의 WAR을 높이게 되면 팀 승률을 높일 수 있음을 의미한다. 이후 WAR과는 팀의 어떤 지표와 연관성이 높은가 확인해보았다. 아래 두개의 그림은 coefficient의 절댓값이 0.3 이상을 가져 WAR과의 연관성이 높은 지표들이다.

|  |  |
| --- | --- |
| Figure 3 타자 WAR과 지표 | Figure 4 투수 WAR과 지표 |

위의 자료들 중 이 프로젝트에서는 투수 데이터 말고 타자 데이터를 기반으로 분석을 진행하였다. 이러한 이유는 투수의 수치는 투수의 능력을 말해주기도 하지만, 수비진과 포수의 개입이 데이터에 영향을 줄 것이라고 생각했기 때문이다. 또한, 타자는 개개인의 능력이 지표 상으로 나타났을 것이라고 예상하였다. 이는 타자의 지표를 통해 분석하는 것이 한 타자의 능력을 잘 표현할 수 있을 것이라 생각했다. 그리하여, 투수의 WAR이 팀 승률에 영향을 많이 미치긴 하지만, 타자의 WAR에 대한 분석을 진행하였다. 타자의 데이터는 kbo로부터 1982년에서 2017년까지 데이터를 Web crawling하였고, 스탯티즈에서는 WAR의 값을 Web crawling하였다. 이때, 타자의 데이터에 추가적인 분석에 쓰일 것이라 생각되는 나이, 연차, 키, 연봉 등에 대한 데이터를 같이 Web crawling 하였다. 이때, 야구에서는 후보 선수나, 규정 타석을 채우지 못한 선수들이 많이 존재하였다. 규정 타석 이상의 선수만을 데이터에서 추출하게 되면, 데이터의 수가 크게 줄었다. 그리하여, WAR에 영향을 줄 수 있는 항목 중에 구장의 변화에 대해서 고려하였다. 적어도 10개의 구장이라 하였을 때, 한 선수가 각 구장에서 적어도 7경기, 3번 이상의 타석을 섰던 선수는 모든 구장에 대해 적응을 하였을 것이라고 생각했다. 그리하여, 타자의 데이터들 중 규정타석의 반 이상의 선수들로 다시 데이터를 만들었다. 이러한 데이터를 기반으로 타자의 WAR과의 연관성이 높은 지표를 찾아보았다.

|  |
| --- |
| Figure 5 타자의 WAR과 지표 |

왼쪽의 그림과 같이 많은 지표들이 타자의 WAR에 연관성이 있다. 이때, 팀의 입장에서 선수를 채용할 때, 어떤 선수가 팀에게 도움이 될 것인지 판단하는 기준이 필요하다. 이때, 한 타자의 높은 WAR은 팀 전체 WAR에 영향을 줄 것이다. 또한, 팀 WAR은 팀 승률에 연관성이 높으므로, 타자의 WAR와 팀 전체의 WAR, 팀 승률 모두와 연관성이 높은 지표를 얻어낼 수 있다. 이러한 지표들이 높은 선수를 위주로 기용을 하게 된다면, 팀의 승률을 올려 더 좋은 팀을 만들어 낼 수 있다. 또한, 지표들을 제시함으로써, 앞서 말한 지표들을 제외하고 다른 데이터들이 낮아 WAR을 저평가 당하는 선수들을 기용할 수 있는 기회를 가질 수 있다. 그리하여, 타자의 WAR, 팀의 WAR, 팀의 승률 세가지 모두에 연관된 지표를 찾고자 하였다.

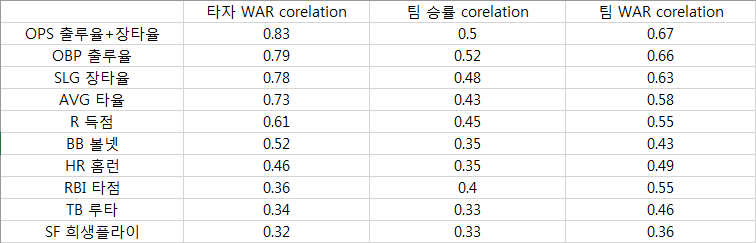


Figure 6 타자 WAR, 팀 승률, 팀 WAR에 연관된 지표들

이 분석을 통하여 위의 지표들을 기반으로 ‘팀은 어떠한 타자를 기용하여 팀의 승률을 높일 수 있을까’에 대한 방안을 제시하였다. 이 방안은 한 선수의 기록을 가지고 살펴보는 것이기 때문에, 앞으로 선수가 어떤 능력치를 가지게 될 것인가에 대한 예측을 하긴 힘들다. 그리하여, 선수가 만약 나이에 따른 WAR 수치가 다르게 나타난다면, 그 선수가 몇 년 후에 어떤 WAR을 가지게 되어, 팀의 승리에 기여할 것인가를 예측할 수 있다. 이러한 생각을 바탕으로 나이에 따른 WAR의 변화를 살펴보았다. 허나, 이 과정에서는 WAR이 과대평가 되었거나, 저평가된 선수들을 제외하여 일반적인 선수들에 한하여 살펴보아야 한다. 이는 예측의 정확성을 높이기 위함이다. 그리하여, 이상치(outlier)를 제거한 후에 데이터를 이용하여 나이에 따른 WAR을 나타냈다.

Figure 7 나이에 따른 타자의 WAR

위의 그래프는 타자의 나이에 따른 WAR 그래프이다. 이는 Aging curve라고도 한다. 추세선은 2차 다항식으로 간단한 다항식임에도 불구하고 R² = 0.7346이라는 높은 결정계수를 가지고 있다. 추세선에 따르면 선수는 24~26살에 최대의 WAR을 가지게 된다. 허나 선수들의 세세한 나이에 따른 WAR의 변화를 알고 싶을 때는 WAR 선을 통해 파악할 수 있다. 위의 결과를 통해서, 한 선수가 미래에 어떤 WAR을 가지게 될 것인가를 예측할 수 있다. 팀의 입장에서 다른 나이, 같은 WAR의 선수를 기용할 때, 미래에 더 높은 WAR을 가질 선수를 기용하여 팀 전체의 WAR을 높임으로써 승률을 더 올릴 수 있게 할 수 있다.

2) 새로운 타격 지표 만들기

WAR은 타자의 능력치를 판별하는 지표이다. 허나, WAR은 매우 복잡한 수식으로 계산되며, 실제로 스탯티즈에서 계산하는 WAR의 식은 공개되어지지 않았다. 만약 WAR 이외에도 선수의 능력치를 판단할 수 있는 지표가 있다면, 그 지표와 WAR을 같이 비교하여 선수의 능력치를 조금 더 자세히 파악할 수 있다. Figure 6을 기반으로 새로운 지표를 구성하게 되면, 타자의 WAR 뿐만 아니라, 팀 승률, 팀 WAR에도 연관성이 있는 지표를 만들 수 있기에, 더 의미 있는 지표가 될 것이다. 이는 새로 만들어진 타격 지표를 이용하여, 팀 승률을 높일 수 있는 타자를 기용할 수 있음을 의미한다.

이러한 분석을 위해서 주성분 분석을 사용하였다. 임의로 랜덤하게 정해진 세 가지의 지표들 간의 주성분 분석을 통하여 새로운 지표를 제작하였다. 이후 새로운 지표와 WAR 간의 연관성이 높다면, 그 지표는 충분히 새로운 타격 지표로의 활용이 가능해 질 것이다. 그리하여 임의의 세 가지 지표로 새로운 타격 지표를 만들어 보았다. New\_value 1은 WAR, OBP, SLG를 통해서 만든 타격 지표이고, new\_value 2는 RBI, R, AVG를 통해서 만든 타격 지표이다.

|  |  |
| --- | --- |
| Figure 8 new\_value1과 WAR | Figure 9 new\_value2와 WAR |
|  |  |

새롭게 만든 두 지표의 RMSE 값을 비교한 결과 new\_value1 이 new\_value2 에 비해 더 좋은 지표인 것을 알게 되었다. 특히 new\_value1은 WAR과 R2값이 0.68로 높은 수치를 보였다. 그리하여 새로운 타격 지표를 만들 수 있었다.

3) 연봉 예측

연봉을 예측하기 위해서, 선수의 능력치 기준인 WAR과 연봉 간의 상관관계를 분석하였다. 그리고 중간 발표 때 질문을 참고하여, 한 선수의 연차가 연봉에 영향을 주는 것이라 생각하게 되었고 연차와 연봉의 상관관계 또한 분석하였다. 이 분석에서 물가 등을 고려하지 않기 위해 2016년의 데이터와 2016년 데이터를 기준으로 한 2017년의 연봉을 비교하였다.

|  |  |
| --- | --- |
| Figure 10 WAR과 연봉 | Figure 11 연차와 연봉 |

위의 R2값을 보아 크게 두 가지 지표 모두 크게 연봉과 상관이 없었다. 허나, 선수의 능력치가 높으면, 연봉을 더 많이 받을 것이고, 연차가 높을수록 선수의 나이가 크게 많지 않다면 노하우와 같은 능력치를 인정받아 연봉을 더 높게 받을 것이라는 예측과 달랐다. 그리하여, WAR과 연차를 이중 회귀 분석을 통해서 두 가지 모두 연관이 있는지에 대해 분석해 보았다. 그 결과는 다음과 같다.

|  |
| --- |
| Figure 12 WAR과 연차의 이중 회귀 분석 결과   * Y = 28886.1278 \* year+9545.3035 \* WAR-1.617 \* 104 |

위의 Y는 이중 회귀 분석을 통하여 얻어낸 식이다. R2값이 이전의 각각에 비해 0.6으로 다소 높은 값이 나왔다. 그래프는 python을 통해 3차원 그래프를 그리는 것에 실패하여 각각 2차원으로 표현한 거이다. 이 결과는 WAR, 연차와 연봉과의 관계가 어느 정도 있다는 것을 의미한다. 그러므로 한 선수의 WAR과 연차를 알게 된다면, 위의 식을 통하여 한 선수의 연봉을 알 수 있게 된다. 만약 이를 앞선 Aging curve와 함께 사용하게 된다면, 한 선수를 데려올 때 더 좋은 선수를 고를 수 있다. 또한, 그 선수의 능력치에 맞는 연봉을 알 수 있어, 그 선수에게 줄 수 있는 최대의 연봉을 미리 책정하여, 과한 소비를 막을 수 있다.

3. 결론

이 프로젝트를 통해 타자의 WAR, 팀의 WAR, 팀의 승률 세 가지 모두 연관성이 있는 데이터를 얻어냄으로써, 한 선수를 기용할 때 선수의 데이터를 통해 팀의 승률에 어떤 영향을 미칠 것인가를 예측할 수 있었다. 또한, Aging curve를 제시함으로써 같은 WAR, 다른 나이의 선수 중 골라야 할 때, 선수의 미래 능력치를 예측하여 팀의 승률을 더 높일 수 있는 선수를 기용할 수 있는 틀을 마련하였다. 이후에 선수를 판단하는 기준인 WAR의 복잡성을 생각하여 주성분 분석을 통하여, 간단하면서도 WAR 만큼의 기준이 될 수 있는 새로운 타격 지표를 만들어냈다. 이로써 한 선수의 능력치를 평가하는 새로운 기준을 만들 수 있었다. 선수의 능력치에 따라 연봉을 예측하기 위해서, 2017년의 연봉 데이터와 2016년의 선수들의 능력치를 비교하였다. 이때, WAR과 연봉, 연차와 연봉은 각각 크게 연관성이 없지만, 두 가지 지표를 이중 회귀 분석을 통해서 만든 지표는 연봉과 어느 정도 연관 있는 값이 제시 되었다. 그리하여, 한 선수의 능력치에 따른 연봉의 상한선을 제시할 수 있는 방법을 제시하였다.

이 프로젝트를 통하여 선수들을 기용할 때, 선수의 능력치가 팀의 승률에 어떤 영향을 끼칠 것인가에 대한 것을 알 수 있었다. 허나, 실제 방대하고 자세한 데이터를 지닌 MLB에서의 분석과 이 프로젝트의 분석을 비교하지 못한 것에 대한 아쉬움이 남았다. 복잡한 수식을 가지고 정확성이 조금 떨어지는 값인 WAR의 정확도를 높이는 작업을 못한 것 또한 아쉬운 점으로 남았다. 하지만 이후에 WAR이 정확해지고, 같은 데이터셋으로 이와 같은 분석을 진행하게 되면, 의미 있는 분석이 진행될 것이라 생각한다. 마지막으로 실제 선수들의 연봉을 예측해보고 싶었으나, 선수들의 정확한 연봉이 아직 공식적으로 발표되지 않아, 비교하지 못한 것 또한 아쉬움이 남았다.