|  |
| --- |
| **1. 주제**  이전 성적 기준, 다음 시즌 타자 OPS 예측 프로그램  **분반, 팀, 학번, 이름**  (가)반, 4팀, 20243259, 강태윤 |

|  |  |
| --- | --- |
| **2. 요약**  이전 시즌과 프리시즌의 성적을 기준으로 여러 변수들을 포함해, 다음 시즌의 타자 OPS를 예측해보는 프로그램 제작이 목표이다.  OPS는 On-base Plus Slugging의 약자로서 야구에서 타자들의 가치를 평가하는 요소 중 하나이며 출루율 + 장타율로 계산한다. 현재엔 타자의 타격능력을 가장 단순 명료하게 설명하지만, 다양한 크고 작은 문제점들이 있다. 여러 문제점들을 다양한 변수들을 통해 해결할 것이다.  이 프로그램을 통해 타자들의 능력치를 다양한 관점에서 볼 수 있을 것이며, 단순히 수치만 보고 판단하는 것이 아닌, 복합적인 요소들이 서로 상호작용하는 지표를 제작할 수 있을 것이다. | **3. 대표 그림**  텍스트, 번호, 폰트이(가) 표시된 사진  자동 생성된 설명  위의 표는 예상OPS를 구할 때, 고려할 수 있는 변수들을 나열한 것이다. 이때의 YOPS는 목적변수로 다음시즌의 예상 OPS로 칭하였다. |

|  |
| --- |
| **4. 서론**  평소 야구에 관심이 많아, 많은 경기들을 챙겨보고, 직관도 갔었다. 야구 경기를 보다 보면, 다양한 수치들이 전광판에 나오는 것을 알 수 있다. 이 수치들에 대해 공부하며 관련된 프로그램에 대해 관심을 키워가던 중, 타자들의 중요한 지표 중 하나인 OPS에 대해 분석해보고 싶어, 이 프로그램 제작을 계획하게 되었다.  OPS란, 야구에서 타자의 공격력을 평가하는 중요한 지표로, 출루율(OBP, On-Base Percentage)과 장타율(SLG, Slugging Percentage)를 더한 값이다. 이 두 지표를 합산하여 타자의 득점 기여도를 종합적으로 평가할 수 있게 해준다. OPS는 단순 타율이나 타점처럼 안타나 홈런의 개수로 파악하는 것이 아닌, 타자가 얼마나 자주 출루 하는 지와 얼마나 강력한 타격을 하는지를 함께 반영한다.  OPS는 출루율(OBP)와 장타율(SLG)를 더함으로 계산한다. 여기서 출루율은 (안타+볼넷+사구)/(타수+볼넷+사구+희생플라이)로 계산한다. 장타율은 (1루타+(2\*2루타)+(3\*3루타)+(4\*홈런))/타수 로 계산할 수 있다. 이렇게 구해진 OPS는 다음과 같은 범위에서 타자의 타격 능력을 평가한다.   * 1.000 이상 : 최상급 타자 (리그 MVP급) * .900 ~ .999 : 매우 뛰어난 타자 (올스타급) * .800 ~ .899 : 우수한 타자 * .700 ~ .799 : 평균 * .600 ~ .699 : 다소 부족한 타자 * .600 이하 : 매우 낮은 타격 성적   이러한 OPS의 장점은 출루율과 장타율 안에 포함되어 있는 다양한 요소를 모두 포함하여 타자의 전체적인 공격력을 종합적으로 평가 가능하다. 또한, 간단한 계산법으로 성적을 직관적으로 평가할 수 있다.  현재의 OPS는 현재 MLB, KBO와 같은 많은 리그에서 사용되는 지표이며, 가장 대표적으로 사용되는 선수들의 평가 기준이다. 다만 각 선수마다 평균치를 기준으로 타율, 장타율, 출루율들이 매 해 변동됨을 알 수 있다. 이는 선수들마다 평균치가 모두 다르며, 각 선수들은 모두 독립적이기 때문이라고 생각할 수 있다. 또한, 출루율과 장타율을 단순 합하는 것이기에, 두 요소의 비중 문제가 있다. 보편적으로는 출루율이 장타율보다 팀 득점과 더 연관이 크다고 알려져 있지만, OPS에는 동일한 가중치로 취급되므로 완벽한 지표가 될 수 없다. 게다가 타자의 클러치 능력(득점권에서의 활약 등)이나 나이, 타격 개인특성, 연도별 기록과 같이 세부적인 상황을 반영하기 힘들다는 단점이 있다.  결국 이 모든 문제점을 해결하기 위해선, 변수 간의 상관관계를 뚜렷하게 파악하고, 나이, 행운의 안타, 타자의 속도, 평균대비 기여율, 누적 연차, 나이와 같은 여러 변수들을 추가 해야 한다. 또한, 단순 1,2,년의 기록이 아닌, 다년간의 기록들을 통해 보다 완벽한 지표가 되도록 보완해야 할 것이다. |
| **5. 본론**  우선 데이터 수집 및 목적 변수를 생성해야 한다. 현재 가장 신빙성 있는 선수들의 기록을 알 수 있는 사이트인 ‘스탯티즈’에서 크롤링을 통해 시즌 별 엔트리에 등록되어 있으며 기록이 존재하는 선수들의 이름, 포지션, 타석 위치, 연도별 성적, 나이 등을 수집해야 한다. (이후부턴 해당 시즌의 성적의 경우 OPS로 칭하고, 목적 변수인 다음 시즌 OPS의 경우, YOPS로 표현할 것이다.) 크롤링을 통해 구한 년도별 성적들을 비교하며 다양한 변수들과 함꼐 YOPS를 예측할 수 있을 것이다.  YOPS를 구하기 위해 필요한 변수들을 유형별로 살펴보면, 독립변수 32개와 목적변수 1개로 볼 수 있다. 이때, 독립변수 중, 타석위치, 포지션은 다음시즌 OPS인 YOPS와 연관성이 크지 않을 것으로 보이기에 제외한다. 남은 30개의 독립변수 중, 그래프를 통해, 가장 핵심적인 변수를 골라내면, 총 4개로 볼 수 있다. 나이, 타수, 홈런 수, war이다. 나이는 유망주들이 언제 성적이 급상승하고, 베테랑 선수들의 성적이 언제 추락하는지를 예측하는데 사용할 수 있다. 타수의 경우, 주전 선수들과 비 주전 선수들의 차이가 YOPS에 영향을 주는지 확인하기 위해서 사용한다. 홈런의 경우, 홈런타자와 단타 위주의 타자들의 차이를 구분하기 위해서 사용한다. 마지막으로 war의 경우, Wins Above Replacement로 대체 선수 대비 승리 기여도라는 뜻으로, 선수들을 비교할 때 유용한 기준이 된다.    위의 그림의 OPS+는 위에서 언급한 주요 변수들을 감안해 만들어진 OPS로 평균치는 100이다. 만약 해당 OPS+의 값이 150이라면, 해당 리그 평균에 비해 50% 더 우수한 선수라는 것이다.  EDA과정에서 확인한 결과, 각 선수들은 선수마다 평균치를 기준으로 타율, 장타율, 출루율등이 모두 독립적이기에 변동된다. 그렇기에 기존에 사용되는 모델이 아닌, 메이저리그에서 사용되는 MARCEL Projection을 활용하여 각 선수의 최근 년도의 기록을 이용하고, 가장 변동성이 적은 위의 4개의 변수를 이용해 가중평균을 이용해 예측한다.  예측 시에는 RandomForestRegressor를 통해 예측을 수행한다. 이는 앙상블 학습 기반의 회귀 모델로, 여러 개의 결정 트리를 사용하여 예측을 수행하는 모델이다. 이는 여러 개의 모델을 조합하여 안정적이며 더 높은 예측 성능과 일반화를 기대할 수 있다. 독립변수들과 전년도들의 기록을 종합해 결론을 추출한다. |

|  |
| --- |
| **6. 결론**    RandomForest Regressor 모형을 구축하고, 이상치 처리 방법 및 다른 년도들의 회귀 모형과 비교 및 이전 시즌들을 통해 이 모형이 합당함을 알아낼 수 있다. 변수 자체적으로 OPS와 상관관계를 통해서, 수치적으로 파악할 수 있기 때문에, 해당 모형에서 나오는 변수 중요도를 통해서 최종 모형에 대한 의미를 해석할 수 있다.  현재는 KBO 선수들의 기록만을 기준으로 YOPS를 예측하기에, 다른 리그의 정보들도 추가해, 활용하면 좋을 것 같다. 또한, 가장 중요한 4개의 독립변수를 중점으로 해석했기에, 향후에는 더욱 다양한 변수들을 추가해 모두 각각의 가중치를 부여하여 해석하면 좋을 것이다. |

**7. 출처**

1. **bluemumin.** (n.d.). *baseball\_ops\_predict*. GitHub. <https://github.com/bluemumin/baseball_ops_predict?tab=readme-ov-file>

2. **나무위키.** (n.d.). *OPS*. <https://namu.wiki/w/OPS>

3. **Dacon.** (n.d.). *[공식 대회 코드 공유 페이지]*. Dacon. <https://dacon.io/competitions/official/62540/codeshare/658?page=1&dtype=recent>

4. **Dacon.** (n.d.). *[공식 대회 코드 공유 페이지]*. Dacon. <https://dacon.io/competitions/official/62540/codeshare/657?page=1&dtype=recent>

5. **Statiz.** (2024). *[타격 통계]*. Statiz. <https://statiz.sporki.com/stats/?m=main&m2=batting&m3=default&so=&ob=&year=2024&sy=&ey=&te=&po=&lt=10100&reg=R&pe=&ds=&de=&we=&hr=&ha=&ct=&st=&vp=&bo=&pt=&pp=&ii=&vc=&um=&oo=&rr=&sc=&bc=&ba=&li=&as=&ae=&pl=&gc=&lr=&pr=50&ph=&hs=&us=&na=&ls=&sf1=&sk1=&sv1=&sf2=&sk2=&sv2=>