제9회「 2021 빅콘테스트」데이터 분석 계획서

* 해당란에 ☑ 표시

참가분야	☐ 이노베이션분야	☑ 데이터분석분야	
세부리그	☐ 루키리그 ☐ 퓨처스리그	☑ 챔피언리그 *데이터분석 분야에 한함	
세부부문 *해당시 체크	☐ 지역활성화 ☐ 중소기업지원 ☐ ECO제주 ☐ 홍수ZERO ☑ 스포츠테크 ☐ 수산Biz		
개인/팀여부	☐ 개인 ☑ 팀(구성원 3 명)	**개인/팀명**	CBO
지도교사명	*루키리그에 한함(선택)		
대표ID	kimjae1119@naver.com		

※ **5장 내외로** 목차는 준수하여 자유롭게 작성

분석 주제명	프로야구 배럴(Barrel)을 통한 타자 성적(OPS) 예측
분석 배경	최근 KBO 리그에 많은 변화가 일어나고 있다. 과거 타율과 타점으로 타자의 능력을 평가하던 것을 최근에는 OPS, WAR(대체 선수 대비 승리 기여도), wRC+(조정 득점 생산력) 등 새로운 지표를 도입하여 평가하고 있다. 예를 들어, 롯데 자이언츠와 NC 다이노스는 메인 전광판에 타자의 타율이 아닌 OPS를 띄우고 있고, 이렇듯 OPS는 이제 야구팬들에게도 친숙한 데이터가 되었다. 이뿐만 아니라 타구-투구 추적시스템 트랙맨 데이터를 수집하여 전력 분석을 진행하고 트래킹 데이터를 중계화면에 등장시켜 팬들의 즐거움을 끌어내고 있다. 이에 트래킹 데이터를 적극 활용하여 KBO 리그만의 배럴 기준을 선정하고 이를 이용하여 타자의 능력을 평가하는 데 사용되는 OPS를 예측하고자 한다.
분석 내용 요약	1. 배럴 타구 정의 기존 MLB에서의 배럴 정의를 KBO만의 타구 속도와 발사 각도의 배럴 타구로 탐색한다. 그리고 득점 생산력이 높은 타구를 배럴로 정의해 분석을 진행한다. 이를 위해 각 타석 상황(누상 주자, 아웃카운트)에서의 기대 득점보다 높은 득점을 기록한 타구를 탐색한다. 2. OPS 앞서 정의한 두 가지의 배럴과 다양한 타격 지표와의 관계를 탐색한다. 그 관계에 기반을 둔 통계적 ML/DL 기법을 통한 예측을 진행한다.
분석방법 및 계획	1. 외부 데이터 수집 KBO 리그 야구 데이터 제공 사이트인 'Statiz'에서 추가적인 데이터 확보를 하고자 한다. 'Statiz 홈페이지 > 기록실 > 시즌기록실 > Playlog'에서 2018년부터 2021년까지 4시즌의 선수별 각 타석 상황 및 결과, LEV, REa, WPs 등의 값을 크롤링한 후 정제하여 사용하려고 한다. 이 데이터는 대회로부터 제공받은

타구 트래킹 데이터에 병합할 계획이다. [출처] http://www.statiz.co.kr/

2. 데이터 전처리

1) 크롤링 데이터에서 동명이인 선수 및 개명한 선수 38명을 식별, 처리하기 위해 제공받은 Player 데이터를 이용해 각 선수에 PCODE를 할당한다.
2) Statiz에서의 수집오류로 타석 상황 데이터가 없는 선수들을 삭제한다.
3) 2018-2021 4시즌동안 경기 후반 지명타자로 출전한 투수들의 타석 기록 3,502개 삭제한다.
4) Statiz 크롤링 데이터의 경우 더블헤더를 구분할 수 있는 기준이 없다. 이를 해결하기 위해 G_ID, PIT_ID로 알 수 있는 경기 날짜, 타구 시각을 얻어내서 크롤링 데이터에서도 더블헤더를 구분할 수 있도록 한다.

3. EDA

1) 타구속도/발사각도별 타구결과 분포를 확인해 본다.
2) 타구 클러스터링 : 배럴타구/준배럴타구/비배럴타구 등으로 나눈 뒤, 그룹별 타구속도와 발사각을 비교해 본다. KMeans Clustring을 진행하여 적합한 타구속도/발사각 기준을 제시해본다.
3) 지표 자체의 상관관계인 자기상관을 확인한다. 자기상관계수가 높을수록 예측력이 높고 이를 변수로 활용하고자 한다.

4. 분석 방법

1) 배럴타구 정의

a. MLB 정의

- 각 타구 속도와 발사 각도의 범위를 설정해 범주형 변수로 설정한다.
- 이 때, Grid Search CV를 이용해 최적의 Hyper Parameter를 선정하여 범위를 적합한 수치로 설정한다.

ex) 타구속도 10km/h 범위, 발사각도 5º 범위

- 각 범위에 해당하는 타구들의 결과가 타율이 0.5이상 장타율이 1.5이상인 범위를 배럴의 조건으로 정의한다.

ex) 타구속도 130~140km/h, 발사각도 20º~25º : 타율 0.6, 장타율 1.7
→ 배럴 O
타구속도 110~120km/h, 발사각도 20º~25º : 타율 0.3, 장타율 1.1
→ 배럴 X

- 배럴의 조건을 만족하는 범위조합을 찾아 타구속도와 발사각도에 적합한 분포를 가정해 연속적인 값으로 변환한다.

b. CBO 정의

- 앞서 크롤링한 데이터로 각 타석의 상황을 추가한다.
- 총 24가지 상황별 타석으로 나눠 평균적으로 몇 점을 추가했는지 계산한다.

ex) 1사1루 상황에서의 기대 득점 : 0.3점

- 상황별 타구의 평균득점도 함께 계산한다.

ex) 1사1루 상황에서 발사각도 30º, 타구속도 140km/h의 타구의 기대득점 : 1.3점

	- 타구의 기대득점이 상황에서의 기대득점보다 높을 경우 배럴로 정의한다. - 배럴의 조건을 만족하는 타구를 찾아 타구속도와 발사각도에 적합한 분포를 가정해 연속적인 값으로 변환한다. ex) 1사1루 상황에서는 타구속도 130km/h이상, 발사각도 30°이상이 배럴 2) OPS 예측 방향 - 목표 : 2021.09.15 ~ 2021.10.08 중 선수 10명의 OPS 예측 - 선수별 새로 정의된 배럴 확률, 나이, 크롤링을 통해 얻은 지표 등을 새로운 변수로 추가하고 OPS, 타율, 장타율과 연관이 있는 변수를 선택한다. - 경기 일정을 홈/원정경기로 나누어 각각 예측한다. - 회귀분석, 머신러닝, 딥러닝 등 다양한 모델을 사용해 성능을 확인한 후 가장 좋은 성능을 가지는 모델을 선택한다. 5. 모델링 1) 회귀분석 모델 - LASSO, Ridge, Elastic Net 등 과적합을 방지할 수 있는 모델들의 성능을 확인해본다. 2) 머신러닝 모델 - 회귀모델에서 뛰어난 성능을 보인다고 알려져있는 RandomForest, XGBoost, SVM 등을 Stacking한 모델의 성능을 확인한다. - 일반적으로 여러 모델을 Stacking할 경우, 개별 모델보다 성능이 향상되기 때문에 좋은 성능을 기대할 수 있다. 3) 딥러닝 모델 - 딥러닝 모델에서는 시간흐름까지 파악할 수 있으므로 선수의 시즌 성적이 아니라 타석 별 결과를 입력 값으로 사용한다. - Multi Layer Perceptron, Recurrent Neural Network(LSTM, GRU)를 사용하여 장기 기억 의존성 문제를 해결하는 시계열 기반의 예측을 하고자 한다.
분석결과 활용 및 시사점	기존 배럴은 타자가 타구를 생산해야만 만들어 질 수 있기에 볼넷, 삼진 등을 고려하지 못한다는 단점이 있다. 그러나 본 팀이 새롭게 정의한 '배럴'은 타석 결과별 가중치를 따로 두어 특정 상황에서의 득점생산력 혹은 승리기여도를 고려할 수 있다. 또한, 이를 타구 생산 외의 결과도 포함하는 타석 당 배럴타구 비율과 함께 사용한다면 타자 성적을 평가하는 좋은 지표가 될 것으로 예상된다. 본 팀이 정의한 '배럴'이 높은 선수가 그 상황에 가장 필요한 타구를 날릴 수 있는 선수라는 점을 활용하면 선발 엔트리나 대타 선수를 구성하는 등 야구 경기를 운영하는 데 매우 효과적일 것이다. OPS는 타자를 평가하는데 가장 합리적이고 경쟁력 있는 지표가 되었다. 이를 이용하여 타자의 가치를 평가함으로써 해당 시즌의 고과를 산정할 수 있고, 미래의 OPS를 예측하여 다음 시즌을 위한 FA선수 영입·트레이드 등 전력 강화 및 팀 운영 계획에 합리적인 근거로 사용될 수 있을 것으로 예상된다.

	최근 WBC, 올림픽 무대에서 한국 야구 대표팀의 성적은 국민들의 기대에 미치지 못했다. 선수 구성, 수비 시프트, 작전 등 데이터를 적극 활용하는 미국, 일본의 벽을 넘지 못한 것이다. 하지만 한국에도 데이터 야구의 열풍이 불고 있는 만큼 야구 데이터 수집 및 분석 체계를 개선하고 본 대회의 결과물과 같은 창의적이고 독자적인 지표를 적극 활용한다면 한국 야구의 수준을 성장시킬 수 있을 것이다. 세이버 메트릭스의 성장은 한국야구가 세계 정상의 자리를 되찾는 출발점이 될 것이다.

※ 제출자료는 최종 출품작 평가시 활용될 수 있음