|  |
| --- |
| **제9회「2021 빅콘테스트」데이터 분석 계획서** |

|  |  |  |  |
| --- | --- | --- | --- |
|  | \* 해당란에 체크 표시 | | |
| **참가분야** | □ 이노베이션분야 V 데이터분석분야 | | |
| **세부분야** | □ 루키리그 □ 퓨처스리그 V 챔피언리그 \*데이터분석분야에 한함 | | |
| **세부부문**  \*해당시 체크 | □ 지역활성화 □ 중소기업지원 □ ECO제주  □ 홍수ZERO V 스포츠테크 □ 수산Biz | | |
| **개인/팀여부** | □ 개인 V 팀(구성원 3 명) | **개인/팀명** | 포세이돈 |
| **지도교사명** | \*루키리그에 한함(선택) | | |
| **대표ID** | duddnd919@naver.com | | |

※ **5장 내외로** 목차는 준수하여 자유롭게 작성

|  |  |
| --- | --- |
| **분석 주제명** | ‘Bayesian Statistics를 이용한 배럴 정의와 OPS예측’ |
| **분석 배경** | 우리나라 야구 리그에는 배럴이 아직 잘 정의되었지 않다. 때문에 이를 데이터를 통해 잘 정의하는 것이 첫 째 목표이고, 선수들의 지난 데이터를 통해 선수들의 2021 후반기 OPS를 예측하는 것이 두 번째 목표이다. 이것이 분석 배경이다. |
| **분석 내용**  **요약** | 장타율 예측: 선수 별 공의 각도와 속도의 사후분포를 Multivariate Normal Bayesian Distribution을 통해서 만든다. 이 사후분포에서 N개의 sample을 추출하고 이 sample을 우리가 MCMC를 이용해 정의한 장타율 산출 모델에 대입한다. 그 결과를 우리의 예측 값으로 사용한다.  출루율 예측: MCMC를 이용해 선수 별 출루율의 사후분포를 만든다. 이 사후분포에서 N개의 sample을 추출해 만든 통계량을 예측 값으로 쓴다.  배럴 정의: 우리나라 선수들의 공의 각도, 속도, 결과 데이터를 이용하여 좋은 공이 될 확률이 높은 각도와, 속도의 범위로 배럴을 정의한다. |
| **분석방법 및**  **계획** | 추가데이터: 무  분석 적용에 활용할 통계분석 기법 방법론: 타구의 결과는 공의 각도와 속도에 의존하지만, 언제나 확률적으로 결정된다. 같은 각도와 속도라 할지라도 각각 다른 결과를 낼 수 있다는 것이다. 그러므로 이러한 ‘확률’을 배럴 정의와 OPS 예측에 녹아낼 필요가 있다고 생각했다. 우리는 그 필요를 충당할 수단으로써 ‘베이지안 사후분포에서의 랜덤추출’을 선택했다. 이 수단이 어떻게 활용되는지 구체적으로 서술하면 다음과 같다.  1. 베이지안 Softmax 회귀를 이용 : 공의 결과는 총 다섯 가지가 있다. 1루타, 2루타, 3루타, 홈런, 그 외 타수에 포함되는 결과가 이에 해당한다.  공의 각도와 속도가 각각 A,B로 주어졌을 때, 안 좋은 공의 결과가 나오는 경우에 대한 1루타가 나올 확률, 2루타가 나올 확률, 3루타가 나올 확률, 홈런이 나올 확률을 베이지안 Softmax 회귀를 이용해 구할 수 있다.    2.    베이지안 다변량 정규 사후분포 이용: 그렇다면 특정 선수가 어떤 각도와 속도를 가진 공을 칠 것인가가 중요한 질문이 된다. 이 질문에 대한 답은 단 하나의 값이 아니다. 선수는 한 달 동안 여러 번의 공을 치게 될 것이고, 우리도 이에 해당하는 개수의 공을 얻어야 한다. 그러므로 특정 선수의 공의 각도와 속도가 어떤 분포를 따른다고 가정하고, 그 분포에서 N개의 (각도,속도)쌍을 추출하고자 한다. 이 때 사용되는 분포가 다변량 정규 사후분포이다. 이 사후분포가 만들어지는 과정은 다음과 같다. 각 선수의 공의 (각도,속도) 평균을 평균으로 가지고, 대각성분이 모두 k인 대각행렬을 공분산 행렬로 가진 다변량 정규분포를 사전분포로 설정한다. (이 때 k는 사후분포가 사전분포에 의존하지 않을 만큼 작은 값으로 설정된다.) 그리고 이 사전분포와 각 선수의 공의 (각도,속도)쌍 데이터를 이용해 다변량 정규 사후분포를 만든다.  3. 베이지안 베르누이 사후분포 이용: OPS의 구성요소인 출루율을 선수별로 예측하는 데에 공의 각도와 속도가 큰 영향을 끼치지 않는다고 판단했다. 공을 치지 않았을 때에도 원활한 출루가 가능하기 때문이다. 따라서 공의 각도와 속도를 출루율을 예측하는 데에 사용하지 않는다. 대신 선수들의 지금까지의 출루율 그 자체를 이용한다. 선수들의 출루율은 0~1의 값을 갖는데, 이를 통해 출루율이 베르누이 분포를 따른다고 생각한다. 그리고 애매모호한 베르누이 분포를 사전분포로 정하고, 선수들 각각의 출루율을 데이터로 이용하여 (출루율을 z/N으로 생각하고, z와 N 모두 이용)선수들 각각의 출루율에 대한 베르누이 사후분포를 만든다. 그곳에서 몇 개의 랜덤샘플링을 한 후, 그것을 평균내어 2021-후반 출루율에 대한 예측값으로 사용한다.  4. 배럴을 정의할 때에는 Softmax회귀와 히트맵,산점도 등 시각자료를 중심적으로 활용할 예정이다. 히트맵을 그릴 때에는 공의 결과별로 다른 가중치를 두어 ‘득점수의 가치’를 배럴 정의에 잘 녹여낼 계획이다. |
| **분석결과 활용 및 시사점** | 만약 이 모델이 선수들의 OPS를 잘 예측하고 그에 따라 배럴을 잘 정의내릴 수 있다면 선수 개개인의 실력을 좀 더 세밀하게 관찰하거나 연구할 수 있다. 그리고 선수를 비교할 때의 축(예측되는 OPS)이 더 생기는 것이기에 야구 팬들의 야구에 대한 재미를 더 가중시킬 수 있다. |

※ 제출자료는 최종 출품작 평가시 활용될 수 있음