|  |
| --- |
| **「2020 빅콘테스트」데이터 분석 계획서** |

|  |  |  |  |
| --- | --- | --- | --- |
|  | \* 해당란에 체크 표시 | | |
| **참가분야** | □ 혁신아이디어분야 ■ 데이터분석분야 | | |
| **세부분야** | ■ 퓨처스리그 □ 챔피언리그 \*데이터분석 분야에 한함 | | |
| **개인/팀여부** | □ 개인 ■ 팀(구성원 5 명) | **개인/팀명** | Coin |
| **대표ID** | djyou128@mju.ac.kr | | |

※ **5장 내외로** 목차는 준수하여 자유롭게 작성

|  |  |
| --- | --- |
| **분석 주제명** | **KBO 팀별 승률, 타율, 평균자책점(방어율) 예측** |
| **분석 배경** | **KBO는 대한민국 최고의 인기 리그이다. 그러나 최근 3년간 관중 수는 하락하고 있으며 올해는 코로나로 인해 무관중으로 리그를 진행하였다. 이러한 관중 하락 이유에는 인기 팀들의 부진과 실책, 볼넷, 오심 등 경기의 재미 감소가 이유로 꼽힌다. 이러한 상황에서 KBO에 대한 분석과 예측을 통해 관중에게는 새로운 재미를 제공하고 팀에게는 승리를 위한 전략 수립에 참고 자료로서 경기 수준 증가에 기여할 수 있다고 판단되었다. 이러한 관점에서 20시즌 KBO 팀별 승률, 타율, 평균 자책점(방어율) 예측을 위한 분석 모델 수립을 진행하고자 한다.** |
| **분석 내용**  **요약** | **(1) EDA**  **- 해당 경기 전 기준 팀/개인별 타율, 평균 자책점 생성**  **- 해당 경기 전 기준 팀별 각종 세이버 메트릭스 지표 생성**  **- 생성한 지표들을 상대 구단별, 최근 기준으로 분할**  **- 생성한 투수지표들을 선발/불펜으로 분할**  **- 시즌별, 구단별 홈/어웨이 승률과 지표 -> 코로나로 인한 무관중이어도 홈 어드밴테지는 여전히 존재함**  **- 선발/불펜의 실점 -> 많은 이닝을 소화하는 선발 못지 않게 불펜 또한 실점에 중요한 요소** |
| **분석방법 및**  **계획** | **분석 방법**  **- 각 구단의 시즌 별 팀의 승률의 순위는 득점 순위와 실점 순위와 거의 동일했다. 이를 바탕으로 결국 팀의 승률은 결국 득점을 많이 하고 실점을 적게 하는 것이라고 생각했다. 따라서 각 데이터들을 종합해 타자의 어떠한 지표가 득점에 영향을 많이 미치는지, 투수의 어떤 지표가 영향을 미치는지를 확인이 필요했고 이러한 과정은 파이썬의 pandas, numpy를 이용해 EDA를 진행했고 시각화가 필요할 때는 파이썬 라이브러리인 plotly와 matplotlib을 번갈아가며 사용했다.**  **- 단순히 안타나 홈런 같은 기존에 있는 데이터와 타율 같은 보편적이 지표로는 득점, 실점과의 상관간계가 높지 않았다.**  **그림(1) 타율과 득점간의 상관관계**    **- 그러나 세이버매트릭스 지표인 OPS와 득점의 상관관계, WHIP과 실점의 상관관계를 구해보니 각각 높은 상관관계를 보였다. 따라서 투수나 타자를 평가할 수 있는 일반적인 지표뿐만 아니라 각종 세이버메트릭스 지표들을 구해 변수로 활용하기로 방향을 정했다.**  **그림(2) OPS와 득점간의 상관관계**    **- 투수 지표와 타수 지표를 나누어 구하였고 변수들을 구하는 방법은 위와 동일한 방법인 파이썬의 pandas, numpy를 이용해 구하였다. 구한 지표들은 다음과 같다.**  **타격지표 – OBP, SLG, OPS, GPA, 타율, BABIP, ISO, HR%, K%, BB%, BBK, wOBA, wRAA, wRC, RC, RC27, XR, SPD, SECA, PSN**  **투수지표 – 평균자책점, RA/9, ERC, FIP, kwERA, LOB, 피홈런율, K%, BB%, K9, BB9, HR9, KBB, whip, P/G, P/IP, P/PA, PFR, CYP, F-E, R-E, OAVG, OOBP, OOSG**  **또한 EDA과정에서 우리가 발견한 특징은 다음과 같았다.**  **1. 상대 구단에 따라서 득점, 실점 뿐만 아닌 각종 지표들은 큰 차이를 나타낸다는 것**  **상대 별로 득점과 상대 별 지표의 상관관계와 상대별 득점과 일반적인 지표의 상관관계를 구했더니 그림(5)와 그림(6)과 같다.**  **그림(3) 상대 별 평균 득점과 일반 OPS의 상관관계**    **그림(4) 상대 별 평균 득점과 상대팀과의 OPS의 상관관계**    **2. 경기를 시작하기 전 최근 경기의 지표에 따라 해당 경기의 지표도 비슷한 결과를 보인다는 것**  **5경기 평균득점에 대한 상관관계는 최근 5경기 OPS가 시즌이 시작하고 해당 경기 전까지의 OPS보다 높은 상관관계를 보였다.**  **그림(5)**    **그림(6)**    **3. 각 구단 별 선발과 불펜의 능력의 차이는 크다는 것.**  **그림(9)와 같이 어떤 구단은 선발이 강하고 어떤 구단은 불펜이 강하다. 또한 그림(10)과 같이 선발과 불펜의 기록에는 차이가 있다. 이닝당 비교를 했을 때 이닝당 출루 허용률인 WHIP이 불펜이 훨씬 높다. 주로 위기상황에 올라오는 불펜의 특성 상 선발과 불펜의 능력은 선발 못지 않게 승패에 영향을 많이 미친다.**  **그림(9)**    **그림(10)**    **따라서 구한 지표들을 상대에 따라서, 최근 5경기의 성적으로만, 투수 지표 같은 경우는 선발/불펜으로 나눠서 따로 구하는 과정을 진행했다.**  **추가데이터 여부**  **- 2020년도 7월 20일부터 제출일 전까지의 기존 주어진 데이터의 형식으로 웹크롤링을 이용하여 데이터를 가져올 예정이다.**  **계획**  **- 변수 추출**  **R의 랜덤포레스트와 단계별 회귀분석, R 제곱 선택법, Mallows의 Cp 선택법을 이용해 알맞은 변수 선택을 진행 중이다.**  **- 딥러닝**  **텐서플로우와 케라스를 통해 LSTM 등 시계열 데이터 예측에 사용되는 딥러닝 모델을 사용하여 최종적으로 앙상블 모델을 구현할 계획이다.** |
| **분석결과 활용 및 시사점** | **- 야구를 좋아하는 팬들에게 승률, 타율, 방어율을 예측해주며 또 다른 즐거움을 줄 수 있고 이는 경기를 직접 보러 가지 못하는 상황 속에서도 프로야구에 대한 흥미도를 높일 수 있는 방안이 될 수 있다.**  **- 분석한 내용을 다양한 관점에서 학습시켜 딥러닝 모델을 만들어보며 최적의 모델을 만들 수 있다.**  **- 구단에 입장에서는 승리를 위한 방법을 제시하며 선수 운용을 좀 더 효율적으로 할 수 있다. 예를 들어 득점에 상관관계가 높은 지표는 OPS 지표이다. 그렇다면 구단은 선수들의 OPS를 올리기 위해 웨이트트레이닝이나 타구 발사 각도를 조절할 수 있고, 해당 지표가 높은 선수들로 타순을 구성할 수 있다.**  **투수 입장에서도 방어율이 낮아도 이닝당 출루 허용률이 높은 투수는 득점권 상황에서 실점할 확률이 높으므로 이에 맞춰 선수를 등판 시킬 수 있다.** |

※ 제출자료는 최종 출품작 평가 시 활용될 수 있음