빅데이터 기반 프로야구 인기도 지표 분석 및 구단별 인기 기여 정도 파악

IT공학전공 2116313 손수경

I. 서론

올해 한국 프로야구(KBO)의 인기가 많이 증가하면서 "역대급 KBO 인기"라는 평가를 받고 있습니다. 하지만 "역대급"이라는 표현이 정확히 어떤 지표에 근거한 것인지에 대한 궁금증이 생겨, 본 연구를 진행하게 되었습니다. 본 프로젝트에서는 2010년대 이후 야구 인기도를 평가할 수 있는 다양한 지표들을 분석하여, 이를 바탕으로 KBO의 대중적 인기를 수치화하고자 합니다. 특히, 인기도를 직접적으로 반영할 수 있다고 판단되는 관중 수를 주요 수치화 지표로 간주하여, 다른 지표들과의 관계를 탐구할 계획입니다. 다양한 지표들을 결합해 구단별 흥행 기여도를 분석할 예정입니다.

II. 분석 목적

본 연구의 주요 목적은 2010년대 이후 한국 프로야구 인기도를 평가할 수 있는 지표들을 탐색하고, 이러한 지표들이 인기도에 미치는 영향을 분석하는 것입니다. 특히 관중 수, SNS 언급량, 유튜브 조회수, Google Trends와 같은 지표들을 통해 KBO의 흥행 요소를 파악하고, 이를 기반으로 구단별로 흥행에 기여한 정도를 분석할 계획입니다.

본 프로젝트의 결과는 다양한 분야에서 유용하게 활용될 수 있습니다. 예를 들어, 미디어 방송사는 구단별 야구 인기도에 미치는 정도를 파악하여 인기 경기의 중계권을 확보하거나, 프로그램 편성을 최적화하는 데 참고할 수 있습니다. 또한 스포츠 마케팅에서는 관중 수와 굿즈 판매율이 높은 구단은 스폰서십 기회를 극대화할 수 있습니다. 구단별로 어느 연령대의 팬이 주로 참여하고 있는지를 분석하여 타겟 마케팅 전략을 수립할 수 있습니다.

III. 데이터 수집

2000년도 이후 야구의 인기도를 파악하기 위한 지표로 연도별 관중 현황, 야구 프로그램 시청률, 야구 관련 기사 개수, 프로야구 굿즈 판매율, 연도별 매진 구장 개수, SNS 언급량, KBO 유튜브 조회수에 대한 데이터를 수집할 예정입니다.

3.1. KBO 공식 사이트를 통한 데이터 수집

연도별 전체 관중 수 및 구단별 관중 수 데이터를 KBO 공식 사이트에서 수집합니다. 구단별 정규시즌 순위 데이터를 수집하여 구단별 정규시즌 순위를 바탕으로 한국 프로야구에 미치는 영향을 분석할 예정입니다.

3.2. 크롤링을 통한 데이터 수집

야구 프로그램 시청률 데이터는 닐슨코리아 사이트에서 크롤링을 통해 시청률 데이터를 수집할 예정입니다. 프로야구 관련 프로그램은 특정 구단과 연관된

방송이므로, 한 경기당 두 팀이 참가하기 때문에 해당 시청률을 두 팀 모두에게 동일하게 적용할 계획입니다. 야구 관련 기사 개수 데이터는 포털 사이트에서 크롤링을 통해 연도별 야구 관련 기사 개수를 수집하여, 미디어 노출이 인기도에 미치는 영향을 평가합니다. 유튜브 조회수 데이터는 KBO 공식 유튜브 채널과 구단별 채널의 조회수 데이터를 크롤링하여, 온라인 상에서의 인기도를 분석합니다.

3.3. API를 통한 데이터 수집

SNS 언급량은 Google Trends 및 트위터 API를 활용하여 야구 관련 키워드 및 해시태그의 언급량 데이터를 수집하고 분석합니다. 이를 통해 팬들의 관심도를 실시간으로 반영할 수 있는 지표를 확보할 수 있습니다.

IV. 야구 인기도에 대한 지표 비교 방법

다양한 지표를 결합하여 야구 인기도와 구단별 흥행 기여도를 평가하기 위한 분석 방법은 다음과 같습니다.

4.1 탐색적 데이터 분석(EDA)

EDA를 통해 수집된 지표의 기초 통계와 분포를 파악합니다. 이를 통해 야구 인기도의 주요 패턴을 사전 탐색할 수 있습니다. 이를 통해 데이터 간 스케일링이 필요한지에 대한 판단을 할 수 있으며 인기도와 각 지표 간의 상관관계 분석이 가능합니다.

4.2 회귀 분석

회귀 분석을 통해 각 지표와 관중 수(인기도) 간의 상관관계를 분석하고, 각 지표가 관중 수에 미치는 영향을 수치상으로 평가합니다. 이를 통해 각 지표의 가중치를 결정하는 데 활용될 수 있습니다.

4.3 AHP 분석

요소 별 상대적 중요도를 쌍대 비교하여 가중치를 산출하는 방법으로, 여러 지표가 있을 때 각 지표의 중요도를 평가합니다. 이를 바탕으로 지표별 가중치를 판단하고, 구단별 인기 기여 정도 파악할 때도 활용될 수 있습니다.

4.4 머신러닝 모델 적용

다중 변수 간의 관계 파악을 위하여 랜덤 포레스트나 XGBoost와 같은 머신러닝 모델을 사용하여 비선형 관계의 상관관계를 파악할 수 있습니다.

V. 프로젝트 예상 결론 및 한계점

이번 프로젝트를 통해, 관중 수 외의 다양한 지표를 결합한 분석이 KBO 인기도 평가에 더욱 정확한 결과를 도출할 수 있음을 확인할 수 있을 것으로 기대됩니다. 또한, 구단별로

어떤 지표가 더 큰 영향을 미치는지 파악하여 구단별 흥행 기여도를 구체적으로 분석할 수 있을 것입니다. 하지만, 데이터 수집 과정에서의 한계도 예상됩니다. SNS 언급량과 유튜브 조회수는 2015년 이후부터 사용률이 급격히 증가하였기 때문에, 이전 데이터와의 비교가 어렵습니다. 또한, 일부 지표들은 정확한 데이터 수집이 제한되거나, 실제 인기도를 완벽하게 반영하지 못하는 상황이 있을 수 있습니다. 특히 시청률 데이터와 기사 개수 데이터의 경우는 기사나 방송의 주관적인 요소가 반영될 수 있습니다. 이러한 한계에도 불구하고, 다양한 지표를 결합한 종합적인 분석은 기존의 관중 수에만 의존하는 평가 방식보다 훨씬 더 확고한 인기도의 지표를 제공할 수 있을 것으로 기대합니다.

VI. 결론

관중 수만으로 야구 인기도를 완벽하게 판단하기에 한계가 있습니다. 관중 수는 과거 데이터를 반영하는 데는 유용하지만, 미래의 인기도를 예측하는 데는 부족함이 있습니다. 따라서, SNS 언급량, 유튜브 조회수, Google Trends 지수와 같은 디지털 지표를 함께 분석하는 것이 중요합니다. 이러한 지표들은 실시간 대중의 관심도를 반영할 수 있기 때문에, 미래의 인기도 변화를 예측하고 마케팅 전략이나 방송 편성에 필요한 준비를 할 수 있는 중요한 정보가 될 것으로 기대합니다.