주성분 분석 및 K-Means 클러스터링을 통한 축구 선수 Scouting system

팀 명 AJ Scout 팀 원 최동훈 지도교수 이 슬 멘 토

개발 동기 및 목적

축구는 데이터 분석이 필수적인 종목이다. 과거부터 유명 구단들은 실력 있는 선수를 영입하기 위해서 거액의 돈을 아끼지 않는다. 영국 BBC에 따르면 EPL(English Premier League) 클럽들이 올해 여름 이적 시장에서 지출한 금액은 19억 파운드(약 3조 원) 정도다. 하지만 잘못된 스카우트(Scouting)로 영입된 선수는 클럽에 막대한 손해를 끼친다.

선수의 경기 형태(playstyle)보다 직감이나 단순 수치(stats)로 선수를 영입하는 경우가 대부분이다. 이러한 영입 방식은 논리성과 객관성을 만족하기에는 어려움이 있다. 수많은 축구 선수가 있고 이들의 능력치를 나타내는 지표들은 많아서 효과적으로 검색하고 처리하여 영입의 선택 기준이 될 필요가 있다.

이 목표를 달성하기 위해 주성분 분석(PCA:Principal Component Analysis), K-평균 클러스터링(K-means clustering), 유사도 분석 등의 데이터 마이닝(DataMining) 알고리즘과 접목하여 선수 영입을 위한 모델/APP를 만든다. 이를 통해 각 구단의 입장에서 핵심 선수가 부상이나 이적 등으로 이탈할 경우 대체자를 영입하는 데 활용할 수 있다. 본 프로젝트에서는 손흥민 선수 (Heung Min Son) 를 예로 들었으나 github에 다른 유명 선수에 대한 추천 선수도 확인할 수 있다.

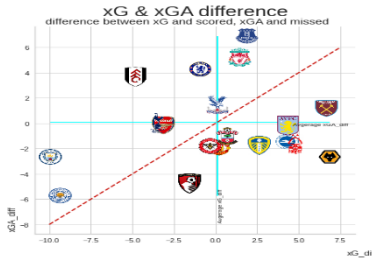
주요기술

1. Web scraping and preprocessing: 첫 번째는 kaggle에 있는 FIFA23 dataset를 활용한다. 필드 플레이어(Non_GK: field player)의 데이터를 가져와 선수마다 능력치를 나타내는 데이터로 재분류한다. 두 번째는 실제 축구에서 사용되는 xG(기대 득점:Expected Goals), xGA(기대 실점:Expected Goals Against) 등의 지표를 가져오기 위해 Understat의 데이터를 가져온다. EPL 팀들의 여러 통계 지표를 가져와 EPL_df라는 이름의 data frame을 만든다.
2. 주성분 분석(PCA): 선수 특성 관련 데이터를 차원 축소를 통해 줄이고자 한다. 알고리즘을 통해 누적 분산량이 전체 중 95%를 넘을 때 기준으로 주성분 개수(n_components)를 4개로 결정하였다.
3. K-means clustering: k를 지정하는 Elbow Method에 따라 5로 설정한 뒤에 선수들을 5개의 군집으로 나누었다.
4. 코사인 유사도 (Cosine similarity): 서로 다른 두 개의 벡터의 유사도를 측정하는 데 특화된 수식으로 클러스터링 연구 분야에서 많이 활용되는 기법이다.

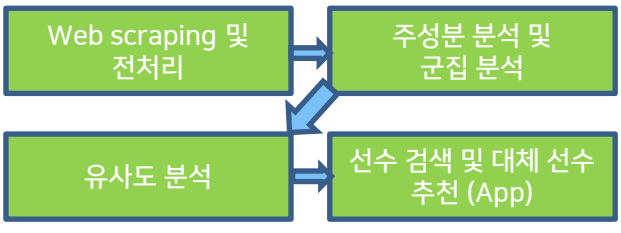
$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

개발내용

Understat.com에서 xG, xGA와 같은 실제 축구에서 사용 중인 데이터를 추출하여 밑의 그림과 같이 시각화하면 구단의 문제점을 파악할 수 있다. (github 링크의 논문 참고)



이에 대한 해결책을 게임 데이터를 통해 분석한다. K-means 클러스터링을 바로 하지 않고 주성분 분석으로 적절히 데이터를 처리한 뒤에 클러스터링을 한다. 이를 바탕으로 유사도 분석까지 하여 선수들의 특성을 파악하고 대체자를 추천해 주는 시스템을 만든다.



결과 및 분석

선수 이름(FullName)을 입력하면 그 선수의 능력치가 방사형 차트를 통해 출력된다. 또한 같은 군집 안에 있는 모든 선수 데이터와의 유사도를 계산하여 그 수치가 높은 순서대로 10명을 출력한다. 예를 들어 손흥민 선수를 입력값으로 할 때 선수와 유사한 능력치 혹은 그래프 모양을 가진 선수들이 similar player로 제대로 출력되는 것을 확인할 수 있다. 이들의 포지션(Position)은 모두 공격수이며 손흥민 선수처럼 빠른 속도를 가진 선수들이어서 충분히 대체가 가능한 선수들이다. 즉, 군집 분석이 제대로 진행되어 비슷한 특성을 가진 선수들이 한 군집 안에 포함되었고 유사도 분석을 통해 대체자를 빠르게 확인할 수 있다.