|  |
| --- |
| **그래프 중심성을 이용한 프로야구 정규시즌 분석**    **요 약**  본 프로젝트에서는 프로야구 정규시즌 분석을 통해 어떤 경기 지표가 승리에 결정적인 요인인지 분석하고 분석 결과가 실제 시즌 팀 순위와 얼마나 일치하는지 확인하기 위해 그래프 중심성을 사용한다.  2022년도 국내 프로야구 정규시즌의 720경기 중 무승부인 12경기를 제외한 708 경기에서, 각 경기와 해당 경기에서 발생한 요인, 경기 결과로 이어지는 그래프를 구성하여 Betweenness centrality, PageRank centrality를 분석한 결과, 득점에 직접적으로 영향을 미치는 타자력과 연관된 요인들이 가장 높은 중심성을 보였고, 실점에 기여하는 투수력과 수비력 순의 중심성 결과를 도출하였다. |

**1. 서론**

**1.1. 연구배경**

기록경기가 아닌 대부분의 스포츠 경기에서는 수비와 공격이 동시에 진행되는 반면, 야구는 공격과 수비가 번갈아 이루어지는데 특히 공격에서는 선수 각각이 상대 팀 투수와 수비 선수들을 상대하게 된다. 따라서 야구에서의 통계자료의 양과 질은 다른 종목에 비하여 더 복잡한 양상을 보인다. 또한 야구는 수비와 공격에서 동료선수들과의 협력 뿐 아니라 선수들의 개별적 활약을 기록하여 통계자료를 만든다. 그만큼 통계자료의 양도 많고 종류도 다양하여 통계분석에 의한 게임 승패의 예측 가능성이 상대적으로 어렵다.

본 프로젝트에서는 이 같은 점에서 흥미를 갖고 프로야구 정규시즌을 분석하려고 한다. 팀 타율, 출루율, 장타율, 실책, era, whip 등의 지표 중에서 어떤 요인이 가장 승패에 영향을 주는지를 그래프 중심성을 활용하여 분석하려 한다. 또한 분석한 결과와 실제로 활용되고 있는 분석방법인 상관계수 측정 및 분류 기계학습 모델을 비교하여 그래프 중심성 결과를 검증해보고자 한다.

**1.2. 연구목표**

경기, 각 요인, 경기결과를 정점으로 구성한 그래프에서 요인별로 기록된 수치를 활용한 가중치를 할당했을 때, 각 요인의 그래프 중심성을 구하고, 타자력, 투수력, 수비력 중 어느 범주에서 높은 중심성의 요인이 많이 검출되었는지 확인하며, 중심성 서열과 상관계수 및 기계학습 모델에서의 변수중요도를 비교하여 그래프 중심성이 해당 범주의 분석에서 어떤 특징을 가지는지 파악한다.

**2. 관련연구**

**2.1. 사회 연결망 분석(SNA)**

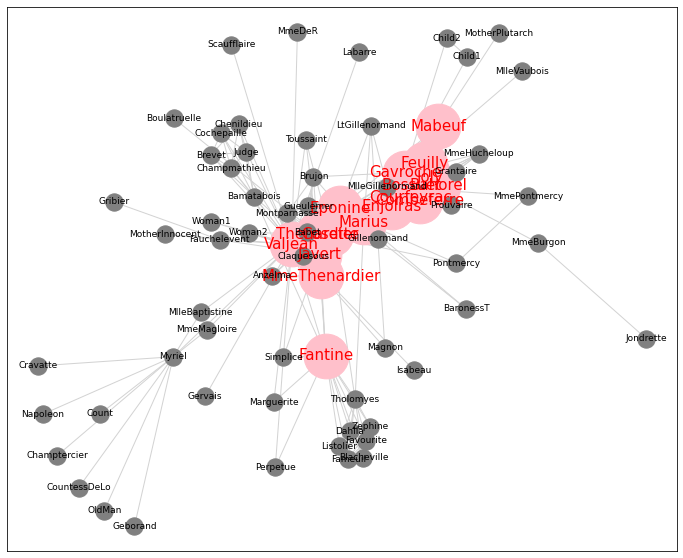
사회 연결망 분석은 개인이나 사회, 집단의 관계를 네트워크의 구조로 분석하는 방법이다. 노드와 엣지를 기반으로 하는 그래프를 사용하여 객체들간의 관계, 각 노드의 중요도를 파악하기 위해 사용한다. 사회 연결망의 분석 기준으로 밀도와 중심성이 있으며, 본 프로젝트에서는 그래프 중심성을 활용한다.

**2.1.1. 그래프 중심성**

네트워크에서 중심성은 하나의 노드가 전체 네트워크에서 어느 정도의 영향력을 가지고 있는지, 즉, 네트워크의 중심에 위치하는 정도를 뜻한다. 사회 연결망을 분석하는데 중심성을 사용하는 경우에는 네트워크에서 연결 정도, 노드의 영향력, 노드 간에 관계가 닿아있는 정도 등, 무엇을 중요하게 볼 것인지에 대한 기준을 정해야 한다. 각 노드의 중심성 수치를 구하기 위한 기준으로는 연결 정도(Degree), 근접(Closeness), 매개(Betweenness), 그리고 고유 벡터(Eigenvector)가 있다.  
 연결정도 중심성은 직접 연결된 노드의 수를 기반으로 측정한다. 이 경우, 하나의 엣지를 통해 직접적으로 연결되어 있는 수가 많을수록 중요도가 올라간다. 근접 중심성은 노드 사이의 최단거리를 기반으로 측정된다. 근접 중심성을 구하면 연결정도 중심성에서 측정할 수 없었던 직접 연결되지 않은 노드와의 관계 또한 수치에 포함시킬 수 있다. 매개 중심성은 다른 노드 사이의 최단거리에 포함된 수를 이용하여 측정한다. 노드들 사이에 위치하는 정도를 나타내고 있으며, 지나는 경로가 많을수록 중심성이 높다. 고유벡터 중심성은 인접행렬의 고윳값에 기반하여 노드에 네트워크에서의 영향력을 계산하고, 연결된 노드와 자신의 영향력을 반영하여 측정한다. 이외에도 고유벡터 중심성에서 파생된 Katz 중심성, 웹 페이지 및 하이퍼링크 등에 활용되는PageRank 중심성, 연결 정도 중심성에서 파생된 가중 연결 정도 중심성 등이 있다[1].

**2.2. Networkx**

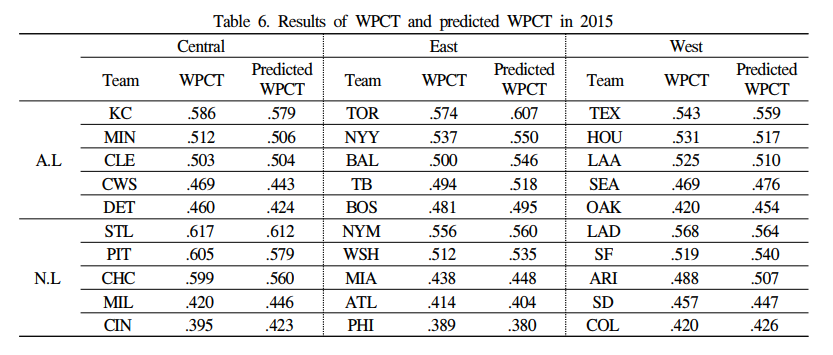
Networkx는 Python에서 그래프와 네트워크를 구현하기 위한 라이브러리이다. Python의 dictionary of dictionary 구조로 그래프를 나타내어 그래프, 방향 그래프, 가중치가 설정된 그래프를 구현할 수 있다. 또한 Matplotlib을 기반으로 그래프의 시각화가 가능하다. 가중치나 직접 설정한 dictionary를 매개변수로 하여 각 정점의 크기, 간선의 색상, heat map 등의 조절이 가능하며, 여러 graph centrality를 도출하기 위한 메소드가 있어 각 정점의 중심성을 파악하는 데에 용이하다.



**[그림 1] Networkx를 이용한 그래프 예시**

**2.3. 기존 통계자료를 이용한 프로야구 분석**

기존의 연구에서는 메이저리그의 경기로부터 79개의 변인을 조사하여, 승률을 반인 변수로, 나머지 78개의 변인을 설명변수로 가지는 다중회귀분석을 수행한다[2]. 전처리 시 반인 변수와 상관관계가 없는 것으로 나타나는 변수와 설명변수 간에 다중공선성이 높게 나타나는 변수들을 제거하고 20개의 설명변수를 사용하여 회귀모형을 구성하였다. 그 결과 각 팀의 승률을 오차수치 3% 이내, RMSE 1.8384라는 높은 예측율을 보였으며, 득점, 홈런, 삼진, 실점, 세이브, 완봉경기 수의 6가지 요인이 승률에 영향을 미치고 있다는 결론을 도출했다.



**[표 1] 이석원, 천영진, “다중회귀분석을 이용한 메이저리그 승률의 모형구축과 예측” 중 승률 및 예측결과 테이블**

**2.4. 기존 연구의 문제점 및 해결 방안**

**2.4.1. 연구의 문제점**

야구 경기에 대한 분석은 이미 치뤄진 경기가 어떤 결과를 초래했는지가 아닌 앞으로 치뤄질 경기가 핵심이 되어야 한다. 해당 연구는 득점, 홈런, 완봉, 실점 등 경기 결과로부터 얻을 수 있는 승리와 연관된 당연한 수치이므로, 승률과의 상관관계가 높고, 예측 성능이 좋다고 할 수 있다. 하지만, 앞으로 이뤄질 두 팀의 경기 결과 예측이나 각 팀이 앞으로 팀의 어떤 부분에 집중하여 성장을 진행해야 할 것인지 등과 같이 연구의 활용안을 생각해보기 어렵다.

또한 해당 연구는 미국 메이저리그로부터 모형을 구성하였다. 메이저리그와 국내의 프로야구는 선수들의 기본적인 수비 기량, 스트라이크 존 크기, 스윙 파워 등 여러 요소에서 차이가 있어 해당 연구를 국내 프로야구에 그대로 적용할 수 없다.

**2.4.2. 해결 방안**

본 프로젝트에서는 국내 프로리그에 대해 각 경기와 요인 및 경기결과를 정점, 팀 기록을 기반으로 한 가중치를 가지고 경기로부터 각 요인을 거쳐 승리와 같은 경기결과로 이어지는 단방향 간선으로 설정한 그래프를 구현한다. 이때, 다른 정점 사이에 위치하는 정도를 나타내는 Betweenness Centrality와 PageRank Centrality를 사용하여 각 요인의 중심성을 구하고, 해당 중심성을 통해 요인들이 경기 결과에 있어 얼마나 큰 중요도를 가지는지 분석한다. 해당 결과로부터 중심성이 높거나 낮은 요인들이 투수, 타자, 수비 기록 중 어느 범주에 속하는지 파악하여 각 팀이 추구해야 할 성장 방향을 정하는 데에 도움을 줄 수 있도록 한다.

**3. 프로젝트 내용**

**3.1. 시나리오 요약**

**3.1.1. 데이터 확보 및 정제**

데이터 확보는 웹 크롤링을 기반으로 한다. 국내 야구 통계 사이트인 Statiz에서 제공하는 선수와 팀의 정보를 기반으로 확보한다. 각 경기의 Box score에 있는 팀 합산정보를 추출해 경기, 요인 사이의 그래프를 구현할 수 있도록 데이터 정제를 수행한다. 정제를 수행하면서, 노드로 사용할 데이터와 관계를 나타낼 데이터를 명확하게 한다.

필요한 요구사항은 다음과 같다.

- 웹사이트의 기록실 및 팀 간에 치뤄진 경기 정보를 스크래핑한다.

- 경기를 치룬 양 팀, 각 팀의 득점, 타자기록, 투수기록, 수비기록 합계를 모두 가져와야 한다.

- 전처리를 통해 적합하지 않은 가중치를 제거하고, 각 요인이 경기에 작용하는 범위를 한정하기 위해 정규화를 수행한다.

**3.1.2. 그래프 구현 및 시각화**

앞서 정제한 데이터를 바탕으로, 그래프를 구현하고, 직관적인 형태로 시각화를 진행한다. 가중치의 부호, 정규화한 수치에 따라 간선의 연결 정도와 간선에 부여된 가중치를 결정하여 그래프를 구현한다. 작업 수행 후 일차적으로 그래프를 시각화 하여 그래프에서 직관적으로 보이는 특징을 확인한다.

필요한 요구사항은 다음과 같다.

- 각 정점과 정점 사이의 간선은 일관성을 가져야 한다.

- 부여한 가중치가 치우치지 않도록 한다.

- 시각화한 그래프는 직관적으로 그래프의 구조를 파악할 수 있도록 한다.

**3.1.3. Graph centrality 도출**

앞서 구현한 그래프로부터 중요변수를 도출하기 위한 중심성 측정 알고리즘을 선정한다. 이후 해당 중심성 측정 척도를 이용하여 각 변수의 중심성을 측정하고, 그 순위를 추출한다. 해당 순위를 통해 직관적으로 볼 수 있는 분석을 수행하고, 중심성과 각 요인이 가지는 상관관계를 파악한다. Networx라이브러리를 활용하여 그래프 중심성을 추출할 계획에 있다. 또한 중심성의 수치를 이용하여 그래프를 2차적으로 시각화 함을 통해 구현한 그래프의 직관성을 높인다.

필요한 요구사항은 다음과 같다.

- 유의미한 결과가 도출될 수 있도록 적절한 중심성 알고리즘을 선정한다.

- 단순한 데이터의 일반화로 인해 생기는 기초적 형태의 결과 도출을 지양한다.

- 그래프 중심성을 구하는 알고리즘은 논리적 근거를 가져야 한다.

**3.2. 데이터 확보 및 정제**

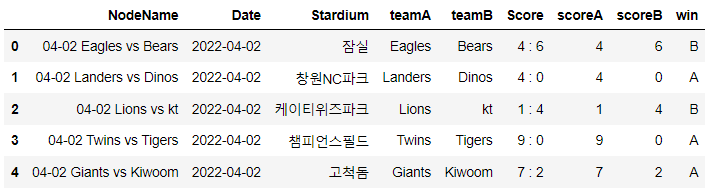
**3.2.1. 데이터 확보**

데이터 확보는 국내 야구 통계사이트인statiz로부터 2022년 프로야구 정규시즌 720경기의 박스스코어를 스크래핑하여 이루어진다. 각 경기의 박스스코어로부터 두 팀의 타자기록, 투수기록, 수비기록의 합계를 추출한다. 총 720개의 경기에서 44개의 요인(타자기록 19, 투수기록 18, 수비기록7)을 크롤링한다. 확보한 데이터의 요인은 다음과 같다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| batter | | pitcher | | defense |
| TPA(총타석) | AB(타석) | IP(이닝수) | TBF(타자수) | IP(이닝수) |
| R(득점) | H(안타) | H(피안타) | R(실점) | PO(풋아웃) |
| HR(홈런) | RBI(타격득점) | ER(자책점) | BB(볼넷) | A(어시스트) |
| BB(볼넷) | HBP(데드볼) | HBP(데드볼) | K(삼진) | E(실책) |
| SO(삼진아웃) | GO(땅볼아웃) | HR(피홈런) |  | \_P(직접실책) |
| FO(뜬공아웃) |  | PIT-S(투구수-스트라이크) | | \_A(간접실책) |
| GDP(병살) |  | GSC(투수평점) | |  |
| AVG(타율) |  | WHIP(이닝당 볼넷안타 허용수) | |  |
| PIT(투구수) | | WPA(승리확률기여도) | |  |
| LOB(잔루) | | RE24(기대득점) | |  |
| LI(현재상황의 중요도) | | LI(현재상황 중요도) | |  |
| RE24(기대득점) | | ERA(평균자책점) | |  |
| OPS(출루율+장타율) | | IR-IS(승계주자-승계주자득점) | |  |
| WPA(승리확률기여도) | | GO-FO(땅볼아웃-뜬공아웃) | |  |

**[표 2] 수집한 요인 및 각 요인 정의**

각 기록은 데이터 프레임에 맞추어 전처리를 시행 후 팀별 기록을 저장한다. 생성된 데이터 프레임에서의 index를 각 경기에 부여된 고유 번호로 하며, 웹 페이지에서 크롤링 시 먼저 언급된 팀을 팀 A, 후술된 팀을 팀B로 구분하여 데이터를 저장한다. 이 때, 각 경기에서 득점을 기준으로 팀 A와 팀 B중 승리팀을 결정하여 경기의 정보를 담고있는 csv파일 또한 함께 저장하도록 한다.



**[그림 2] 추출한 경기 정보의 일부**

**3.2.2. 데이터 정제**

수집한 데이터를 그래프로 구현하기 위해 5가지 전처리를 수행하였다. 전처리를 수행한 후 각 팀의 기록(타자, 투수, 수비기록), 경기 정보를 하나의 데이터 프레임에 담는 작업을 진행했다. 수행한 전처리 과정은 다음과 같다.

**3.2.2.1. 무승부 경기 제거**

구현하고자 하는 그래프는 경기 → 요인 → 경기 결과 와 같은 형태의 그래프이므로, 무승부 경기에서 승리와 패배를 결정짓는 요인을 객관적으로 판단하기 어려워 720개 경기중 무승부 12경기를 제거하였다.

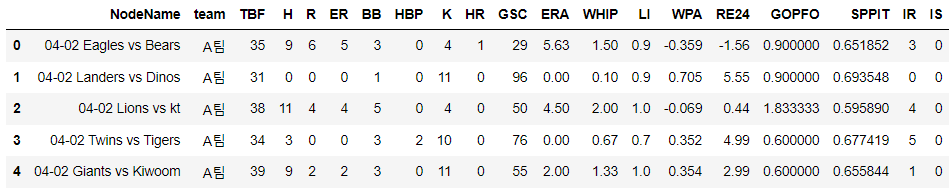
**3.2.2.2. 투수기록의 비율형 데이터 연속형 데이터로 변환**

투수기록의 GO-FO (땅볼아웃-뜬공아웃), PIT-S (피팅 수-스트라이크), IR-IS (승계주자-승계주자 득점) column은 float-float 형식으로 가중치를 부여하기 위한 숫자로 계산할 수 없는 상태였기 때문에 각 수치를 적절한 형태로 변환했다. 변환 방식은 다음과 같다.

GO-FO : 뜬공과 땅볼의 비중을 비교하여 투수의 실력을 측정하는 요소중 하나인 GO/FO라는 통계가 존재하므로 GO를 FO로 나누어 변환. 그래프 내에서는 GOPFO.

PIT-S : 피팅 수 중 스트라이크의 비율으로 투수의 실력을 나타낼 수 있을 것이라 판단하여 S/PIT로 변환. 그래프 내에서는 SPPIT.

IR-IS : 서로 별다른 상관관계를 보이지 않아 IR과 IS를 분리.



**[그림 3] 변환한 투수기록(A팀의 기록 일부)**

**3.2.2.3. 경기 단위로 팀 기록 합산**

경기 단위의 그래프를 구성하기 위해 하나의 경기에 포함된 두 팀의 기록을 합산한다. 가중치 계산을 위해 두 팀 중 승자팀의 기록에서 패자팀의 기록을 뺀 값을 합산된 데이터 프레임의 각 열에 할당한다. 이때, 타자와 투수의 기록에 동일한 부분이 있었는데, 타자입장에서의 득점, 투수입장에서의 실점처럼 서로 상반되지만 동일한 특성이 존재했다. 이에 해당되는 요인은 R(득점), H(안타), HR(홈런), BB(볼넷), HBP(데드볼), LI(승부의 중요한 순간에 타석에 선 정도), RE24(기대 득점)이 있고, 해당 기록들은 각 기록에 직접적으로 관여하는 범주의 기록을 사용하는 방식으로 통합하였다. R, H, HR, LI, RE24는 타자의 배팅 및 진루를 통해 결정되는 요인이므로 타자의 기록을 사용하였고, BB, HBP는 공을 던지는 투수에 의해 결정되는 상황이므로 투수의 기록을 사용하였다.

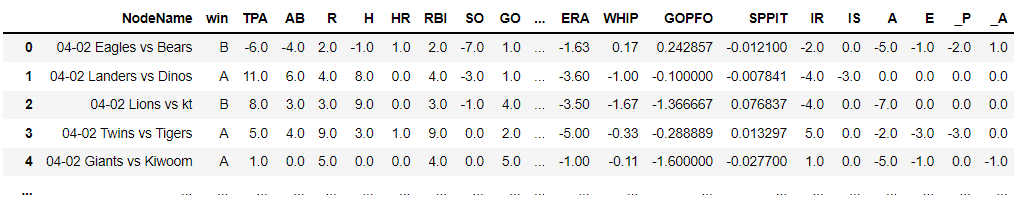
**3.2.2.4. 경기와 관련없는 column 제거**

수비기록에 존재하던 PO (경기 총 아웃 수), IP (총 이닝 수), GDP (병살)은 각각 24 혹은 27(PO), 81(IP), 0(GDP)으로 경기의 내용과 상관없이 일정한 값을 가지기 때문에 제거했다.

**3.2.2.5. 각 수치 정규화**

각 요인이 경기에 작용하는 범위를 동일하게 하기 위해 가중치로 쓰여질 수치를 정규화한다. 합산된 각 기록에 대해 min-max 정규화를 수행하여 모든 요인의 수치를 0-1구간에 위치하도록 한다.

해당 전처리과정을 통해 데이터를 708개의 경기, 32개의 요인으로 구성된 테이블과 각 간선에 할당할 distance 값을 가지는 정규화 테이블으로 정제했다.



**[그림 4] 전처리 결과 테이블**

**3.3. 그래프 구현 및 시각화**

**3.3.1. 그래프 구현**

그래프는 두 개의 전처리 결과 테이블로부터 다음과 같은 과정을 거쳐 구현한다.

1) 데이터 프레임의 각 인덱스를 노드(경기 번호)로 지정

2) 각 열의 이름을 주요 변수로써 그래프에 노드로 추가

3) 경기 결과 노드 (‘win’) 그래프 추가

4) 경기 노드에 변수의 수치가 양수인 경우 변수를 정규화한 수치를 가중치로 가지는 단방향 엣지 추가

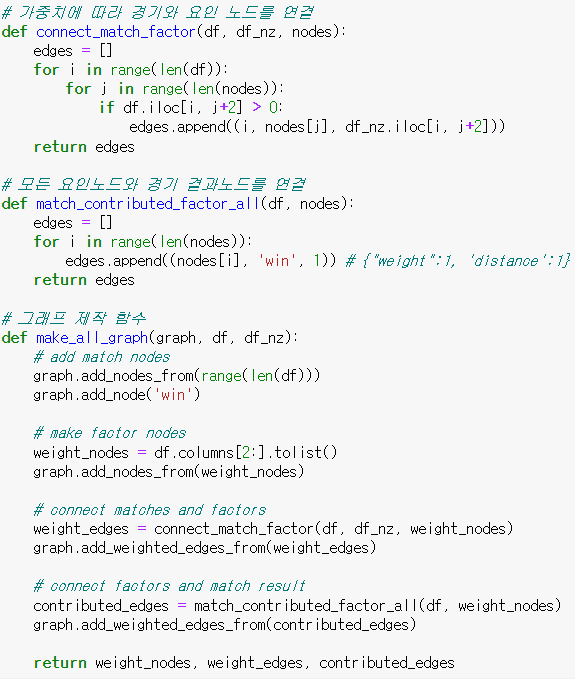
5) 수치가 음수인 경우 해당 경기 노드는 변수로 향하는 엣지를 가지지 않음

6) 각 변수 노드는 동일한 가중치(1)를 가지는 경기 결과 노드로 향하는 단방향 엣지를 포함

이 과정을 거친 그래프는 708개의 경기, 32개의 요인, 1개의 승리결과의 741개의 정점과 각 경기에서 (승리팀의 요인) – (패배팀의 요인) 값이 양수인 경우 해당 요인 정점과 정규화된 가중치를 가지고 연결된 간선 10,162개, 각 요인과 승리 결과 정점을 연결하는 32개의 간선을 합하여 10,194개의 간선으로 구성된다.

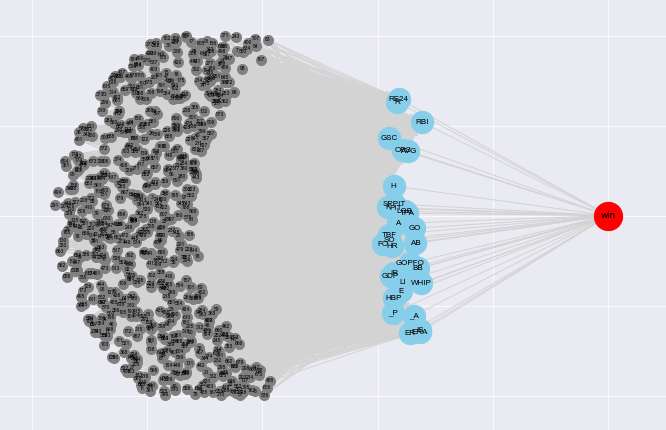
해당 그래프에서 사용되는 가중치는 PageRank centrality를 계산할 때만 사용되는 가중치로, Betweenness centrality를 계산할 때 사용하지 않는다. 그러므로 해당 가중치는 해당 경로를 통해 이동하는 비용을 나타내는 것이 아닌 PageRank에서의 방문 빈도, 즉 해당 경로의 중요도를 나타내고 있다.

그래프 구현에 필요한 코드는 모두 함수형 구현하여 추후 다른 데이터로 그래프를 생성하거나 활용하기 용이하게 함수형으로 구성하였다. 그래프 구현에 사용된 코드는 다음과 같다.

**[그림 5] 그래프 구현 코드**

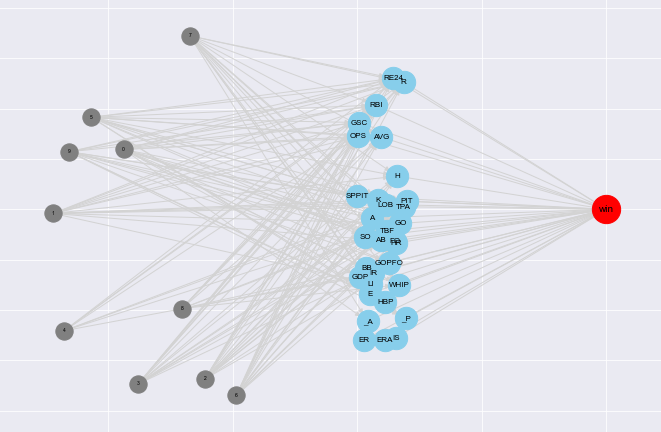
**3.3.2. 그래프 시각화**

시각화에는 networkx, matplotlib.pyplot 라이브러리를 사용하였다. 그래프가 한방향으로만 진행되는 소규모의 그래프이므로, random 라이브러리를 활용하여 경기 노드를 왼쪽에 배치하고, 중간에 있는 요인을 거쳐 경기결과(승리)에 도달할 수 있도록 그래프를 시각화하였다. 각 경기 노드에는 회색, 요인 노드에는 하늘색, 경기결과 노드에는 빨간색, 간선에는 밝은 회색의 색상을 부여하여 그래프의 각 구성요소가 구분되어 보일 수 있도록 시각화 했고, 그래프 구현 시에 networkx라이브러리의 DiGraph라는 방향그래프 객체로 그래프를 구현하여 시각화 할 때 그래프 상에 방향 또한 표시된다. 전체 그래프를 시각화한 결과는 다음과 같다.



**[그림 6] 모든 정점 및 간선을 시각화한 그래프**

해당 그래프에서는 경기의 노드가 708개이고, 앞서 언급한 경기의 노드와 각 요인을 연결하는 10,162개의 간선을 모두 구현하기 때문에, 간선이 중첩되어 경기와 요인 노드 사이가 간선의 색으로 칠해진 형태를 띄며, 요인이 가지는 연결정도의 많고 적음을 확인할 수 없다. 또한 방향이나 가중치의 정도도 확인하기 어렵다.

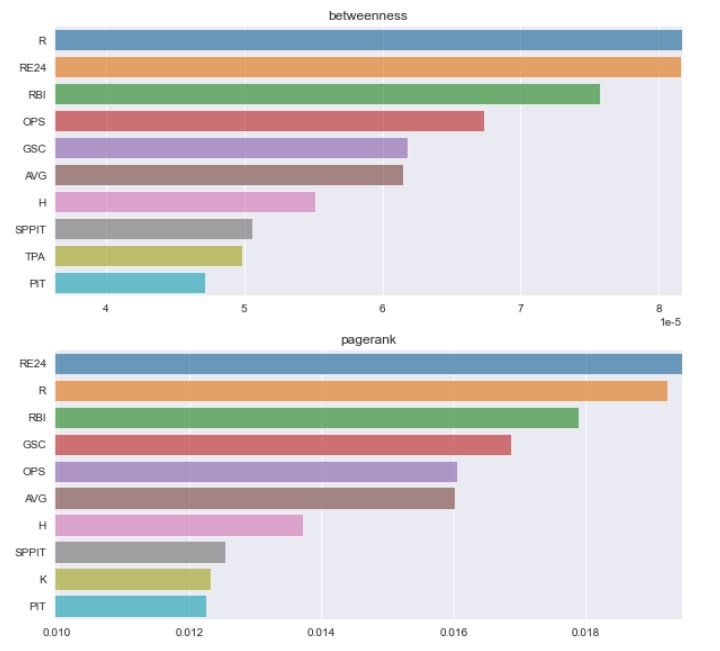


**[그림 7] 0-9번 경기를 시각화한 그래프**

앞선 전체 그래프에서 식별에 어려움이 있어 적은 수의 경기 노드를 포함하는 그래프를 재구성하여 시각화 하였다. 해당 그래프에서는 각 요인의 연결 정도를 확인할 수 있으며, 승리와 높은 연관을 보일 것으로 추정되는 R(득점), AVG(타율), OPS(장타율), RE24(기대득점) 등의 요인에 많은 수의 간선이, 승리하는 팀에서 밀리지 않을 것으로 추정되는 ER(실책), ERA(평균 실책률), IS(승계주자 득점) 등의 요인에 없거나 적은 수의 간선이 연결되어 있는 것을 확인하였다.

**3.4. Graph Centrality 도출**

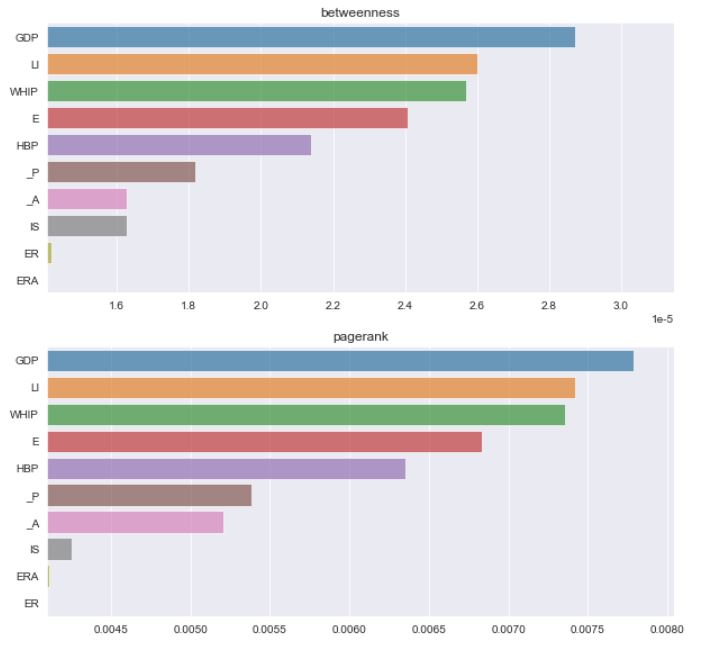
networkx의 내장 메소드를 통해 각 요인에 대한 Betweenness centrality, PageRank centrality를 도출했다. 각 중심성에 대한 상위 10개 요인은 다음과 같다.



**[그림 8] 중심성 상위 10개 요인**

두 중심성에서 R, RE24, RBI, OPS의 타자기록과 관련된 변수가 최상위 중심성에 위치하고 있고, 상위 10개 기록 중 Betweenness 중심성의 경우 8개, PageRank 중심성의 경우 7개의 값을 타자기록이 차지하고 있다. 수비기록은 상위 10개의 요인에 등장하지 않았지만, 투수기록 중 GSC는 타율(AVG)보다 높은 위치에 모습을 보이며, SPPIT(투구 중 스트라이크 비율) 또한 높은 중심성을 보인다.

홈런은 중심성 수치가 상위 10개 요인에 포함되지 않는 비교적 낮은 값을 가지는데, 이는 매 경기마다 홈런이 발생하지 않기 때문에, 그래프에서의 중요도가 다소 떨어지는 경향을 가지는 것으로 추측된다.



**[그림 9] 중심성 하위 10개 요인**

다음은 승리로 이어지는 그래프에서 중심성 수치가 요인들 중 하위 10개에 속하는 요인들의 중심성을 그래프로 나타낸 모습이다. 하위 10개 중심성의 경우 두 중심성이 동일한 순위를 보인다. ER, ERA가 가장 작은 중심성을 보이고 있으며, 2개의 타자기록, 3개의 수비기록, 5개의 투수기록이 그래프 상에 나타난다.

이를 통해 해당 그래프에서 그래프 중심성을 통해 도출된 값은 승리와의 연관성을 나타냄을 알 수 있다. 일반적으로 많은 경기와 연결될수록 높은 중심성을 추출할 수 있다. 야구 경기에서 중요한 요인일 것으로 추측할 수 있는 대표적인 요인인 홈런과 완봉경기가 해당 그래프의 중심성에서 상위권을 보여주지 못한 이유도 발생한 경우 승리로 이어질 확률이 매우 높지만, 모든 경기에서 발생하지 않기 때문에, 승리한 경기에서 일반적으로 높은 중심성을 보이는 요인들보다 낮은 값을 보인다고 예상하고 있다.

**4. 결과 분석**

**4.1. 검증**

해당 그래프 중심성 수치를 통해 도출된 결과의 신뢰성을 실제로 예측이나 상관관계를 구하는데 사용되는 머신러닝 모델, 피어슨 상관계수를 이용하여 비교 검증한다.

**4.1.1. 전처리**

기존의 데이터 프레임에서와 같이 승자팀의 기록에서 패자팀의 기록을 빼지 않고, 먼저 언급된 A팀이 경기에서 승리했는지를 기준으로 label을 설정한다. 각 feature는 A팀의 기록에서 B팀의 기록을 뺀 값을 저장하며, label은 A팀 승리 시 1, 패배 시 0으로 할당한다.

**4.1.2. 평가 지표 도출**

그래프 중심성으로 도출된 서열과 값을 평가할 지표를 선정하고 그 값을 추출한다.

**4.1.2.1. 피어슨 상관계수**

피어슨 상관계수는 두 변수간의 선형 상관관계를 계량화한 수치 이며, 데이터의 label은 승리와 패배의 반대되는 요소로 형성되는 이진 분류이기 때문에, 피어슨 상관계수를 추출하여 유의미한 값을 확인할 수 있을 것으로 판단했다.



label과 각 요인간의 상관계수를 구한 결과는 다음과 같다.

해당 결과에서 0-1구간에 할당되어야 할 상관계수가 0에서 약 0.5에 한정되는데, 이는label이 0, 1을 기준으로 설정되어 음수값을 가지지 않는 것이 원인이다. 서열의 변화는 다소 보이나, 각 수치는 그래프 중심성을 통해 도출한 값과 유사한 경향을 보인다.

**[그림 10] 상위 10개 피어슨 상관계수**

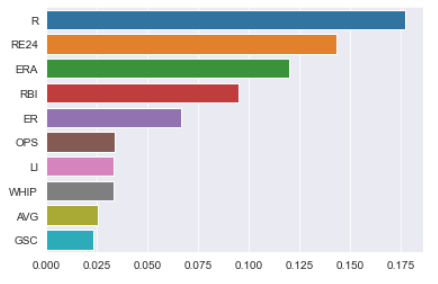
**4.1.2.2. 기계학습 모델의 변수 중요도**

일반적으로 이진 분류에 사용되는 대표적인 기계학습 모델 4가지(로지스틱 회귀, 서포트 벡터 머신, XGBoost, Random forest)중 모든 feature를 사용하여 학습한 모델 중 정확도와 f1-score가 가장 높은 모델의 feature importance와 비교한다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **LR** | **SVM** | **XGB** | **RF** |
| Accuracy | 0.8028 | 0.8380 | 0.9154 | 0.9225 |
| F1-score | 0.8028 | 0.8413 | 0.9142 | 0.9208 |

**[표 3] 분류 모델의 정확도 및 F1-score**

각 모델의 예측 결과는 다음과 같고, random state, 초매개변수를 수정해가며 교차검증을 진행했을 때, Random forest 모델의 성능이 일반적으로 가장 우수하여, 해당 모델의 변수 중요도를 추출해 평가지표로 사용한다.



**[그림 11] Random forest 모델 변수 중요도**

변수 중요도를 추출한 결과는 다음과 같다. 변수 중요도에서 R과 RE24의 수치 차이가 그래프 중심성이나 상관계수에서보다 크게 나타나는 이유는 두 요인이 강한 상관관계를 가져 다중공선성이 높게 나타나 불순도를 더 확실하게 줄일 수 있는 R이 결정적인 요소로 분류를 더 많이 수행했기 때문으로 추측된다.

**4.3. 비교분석**

**4.3.1. 그래프 중심성과 상관계수**

상관계수가 높게 측정된 요인은 경기의 승리, 즉 점수에 영향을 미치는 요인들이 높게 측정되었다. 그 중에서도 득점과 직접적으로 연관되는 출루율이나 장타율, 평균 타율의 순위가 그래프 중심성에서와 달리 득점이나 기대득점보다 높은 위치에 있으며, 홈런과 같이 경기를 결정짓는 요인도 상관계수 9위에 위치해 있다. 투수력이나 수비력을 나타내는 요인들은 야구 경기에 중요한 것은 맞으나, 득점보다는 실점에 연관되는 요인들로 비교적 낮은 순위를 차지하고 있다.

중심성의 경우 투수력, 투구 중 스트라이크 비율 등 투수의 기록 또한 상위 요인에 위치하고 있다. 득점을 위주로 각 수치를 계산하는 상관계수에서 보다 경기 승리와의 연관성을 더 높게 확인하고 있음을 알 수 있다. 하지만, 자주 등장하지 않아도 경기에 중요하게 작용하는 홈런과 같은 결정적인 요인을 확인할 수는 없었다.

**4.3.2. 그래프 중심성과 변수 중요도**

학습된 기계학습 모델에서의 변수 중요도에는 실책이나 평균 실책률, 볼넷 안타 허용 수와 같이 그래프 중심성의 최하위에 위치한 변수들의 중요도가 높게 측정되었다. 이는 Random forest 모델이 예측률을 높이기 위해 불순도를 낮출 때, 패배에 기여하는 요인 또한 이용하였고, 그 요인들을 포함하여 예측에 사용한 비중을 서열로 나타낸 것임을 나타낸다.

그래프에서 경기 결과를 패배로 하여 가중치 및 각 요인 별 수치를 패자팀의 기록 – 승자팀의 기록을 하는 방식으로 개선하여 구현한다면, 패배에 기여하는 요인 또한 알 수 있을 것이지만, 승리에 기여하는 요인과 패배에 기여하는 요인 중 경기에 어떤 요인이 더 결정적으로 작용했는지를 수치적으로 알기는 어렵다. 이는 승리와 패배 결과 모두 나타낼 수 있는 그래프의 구현, 각 가중치가 이상치에 영향을 덜 받도록 하는 새로운 전처리 방법 등을 사용하여 개선할 수 있을 것으로 예상된다.

**4.4. 결과 분석**

구현된 그래프에서 측정된 각 요인의 중심성 서열은 다음 세 가지의 특징을 가진다.

* **해당 그래프에서의 그래프 중심성 서열으로는 승리하는 팀이 일반적으로 포함하는 변수의 서열을 알 수 있다.** 매 경기마다 등장하지 않는 홈런과 같은 변수는 상위권에 위치하지 않는다.
* **그래프 중심성을 통해 보았을 때, 일반적으로 경기의 승패에 연관있는 기록은 타자 기록에 속하는 경우가 가장 많았고, 그 다음으로 투수 기록, 수비 기록 순으로 나타났다.** 중심성 상위 10개 기록에서 7개가 타자 기록, 3개가 투수 기록, 하위 10개 기록에서 2개가 타자 기록, 5개가 투수 기록, 3개 수비 기록이 존재하며, 개수가 아닌 비율로 볼 때, 수비 기록 중 절반이상이 하위 10개 기록에 속하며, 승리와의 연관이 저조하나, 패배와 연관되어 있을 것으로 판단된다.
* **중심성을 반대로 구하여 각 변수가 패배에 기여한 정도를 구할 수 있으나, 두 중심성 수치를 결합하여 하나로 나타내는 것에는 어려움이 있다.** 머신러닝 결과와 같이 경기 결과를 통합하여 표현하기 어려운 구조의 그래프로 한 쪽의 그래프를 구할 수는 있지만, 두 그래프를 구하여 중심성을 구했을 때, 경기에서 변수의 중요도를 하나의 서열로 나타내기 어렵다.

**5. 결론**

**5.1. 결론**

본 연구에서는 프로야구 정규시즌의 여러 기록과 경기, 승리 사이의 기록들에서 중요도를 파악하고자 네트워크를 구성하고 그래프 중심성을 구하였다. 2022년의 각 경기와 수집 후 전처리를 거친 32개의 기록을 정점으로, 각각의 기록에서 승자팀과 패자팀의 차를 구하고 정규화한 값을 가중치로 활용하여 간선을 만든 그래프에서 Betweenness Centrality와 PageRank Centrality를 사용하여 중심성을 측정하였고, 승리한 경기와 연관이 높은 요인들을 확인할 수 있었다.

또한 실제 분석에 사용되는 기계학습 모델이나 상관계수를 이용하여 그래프 중심성 수치로 추출된 각 요인 서열의 신뢰성 검증을 수행했다. 구현한 그래프에서 각 요인의 중심성 서열을 통해 각 요인이 경기 결과와 연관된 정도를 파악할 수는 있었으나, 패배에 기여하는 요인과 승리에 기여하는 요인 중 어느 요인이 더 중요한지는 파악하기 어렵다는 결함도 존재했다.

**5.2. 추후 연구 방향**

향후에는 경기 패배로 이어지는 그래프를 만들고, 해당 그래프에서의 중심성을 측정하여 패배에 기여한 요인, 드물게 등장해 평가할 수 없는 요인을 분류하고자 한다. 또한 해당 결과값을 활용해 각 팀의 기록을 보고 팀의 강점, 약점을 분석해 보았을 때, 어느 부분을 보완하고 준비할 때 승리 확률이 높아지는지 알 수 있도록 예상하는 연구를 진행하여 프로젝트를 발전시킬 수 있을 것이다.

**6. 참고문헌**

[1] 조태수, 한치근, 이상훈, “그래프 중심성들을 이용한 그래프 유사도 측정”, 한국컴퓨터정보학회논문지, 2018.

[2] 이석원, 천영진, “다중회귀분석을 이용한 메이저리그 승률의 모형구축과 예측”, 한국자료분석학회, 2017.

[3] Andrew Disney “PageRank centrality & EigenCentrality”,

<https://cambridge-intelligence.com/eigencentrality-pagerank/>, 2022.

[4] networkx documentation <https://networkx.org/documentation/stable/index.html>