인공지능 기반 팀 스포츠 시즌 승점 회귀 및 순위 산출 모델 연구

이승주, 이상민, 김황남*
고려대학교 전기전자공학부
{joon8958, lsm5505, hnkim}@korea.ac.kr

AI-Driven Model for Season Points Prediction and Ranking in Team Sports

요 약

팀 중심 스포츠 리그의 시즌 최종 테이블(승점·순위) 예측은 전력 변동과 외생 변수의 누적 효과로 인해 난이도가 높다. 본 연구는 선수 단위 성과 지표와 이적시장 및 팀 수준 지표를 통합하여 시즌 최종 결과를 추정하는 일반화된 예측 프레임워크를 제시한다. 제안 모델은 공유 다층 퍼셉트론(Multi-Layer Perceptron, MLP)로 선수 표현을 학습하고, 분가중 평균과 prior Top-K를 활용한 순서 불변 집합 풀링, 그리고 EMA - Last - Delta 기반의 단순 시계열 요약을 통해 시즌 수준 임베딩을 구성한다. 이 임베딩을 팀-시즌 스칼라와 결합하여 승점을 회귀하고, 예측 승점을 내림차순 정렬해 최종 순위를 산출한다. 이러한 설계는 데이터 규모가 제한적인 환경에서도 안정적으로 학습되며, 핵심 선수군과 스쿼드 뎁스의 상대적 기여를 해석 가능하게 반영한다. 본 연구진은 제안된 모델의 학습 및 성능 검증을 위해 워크-포워드 설정(대상 시즌 직전 3개 시즌 학습)을 채택하고, 프리미어리그 데이터를 사례로 2015/16 - 2024/25 기간을 구축하여 7개 시즌을 평가하였다. 예측 순위와 실제 순위 간 평균 제곱근 오차(Root Mean Squared Error, RMSE)는 평균 4.157, 최소 2.864를 기록하였다. 결과적으로, 선수·이적·팀 정보를 결합한 집합 기반 접근이 팀 기반 스포츠 리그의 시즌 최종 테이블을 유의미한 정확도로 사전 추정할 수 있음을 확인하였다.

Ⅰ. 서 론

팀 중심 스포츠 리그의 시즌 최종 테이블(승점 및 순위) 예측은 선수단 구성의 변동, 재정, 이적시장 요인, 스케줄, 부상 등 외생 변수의 누적 효과로 인해 높은 난이도를 갖는다. 본 연구는 선수 단위 성과 지표와 이적시장과 팀 수준 지표를 통합하여 시즌 최종 결과를 추정하는 인공지능 기반 예측 프레임워크를 제시하며, 잉글랜드 프리미어리그(PL) 데이터를 사례로 그 타당성을 검증한다. 제안 모델은 선수 표현을 공통 공간에서 학습하고, 순서 불변 집합 요약 및 시계열 요약을 통해 시즌 수준 임베딩을 구성한 뒤, 팀-시즌 스칼라와 결합해 승점을 회귀하고 예측 승점의 내림차순 정렬로 최종 순위를 산출한다. 이 설계는 데이터가 제한된 환경에서도 안정적인 학습을 유도하고, 핵심 선수군과 스쿼드 뎁스의 상대적 기여를 해석 가능하게 반영하며, 불필요한 튜닝 부담을 낮춘다 [1]. 본 연구진은 워크-포워드 설정으로 학습·검증을 수행하여 실제 운영 환경에 부합하는 평가 체계를 마련하였다.

Ⅱ. 본론

1. 데이터 수집

본 연구진은 제안된 모델의 학습 및 성능 검증을 위해 사례 데이터로 프리미어리그(PL) 2015/16 - 2024/25의 10개 시즌을 구축하였다. 선수 성과 지표는 Understat에서 제공하는 시즌별 90분 환산 기록을 활용하였다. 핵심 입력은 (1) 기대득점(90분 기준), (2) 비페널티 기대득점(90분 기준), (3) 기대도움(90분 기준), (4) 공격 전개 기여도(90분 기준), (5) 빌드업 기여도(90분 기준), (6) 출전 시간, (7) 포지션 범주(GK/DF/MF/FW), (8) 직전 시즌 대비 동일 포지션 내 성과 백분위, (9) 전년도 로스터 부재 여부(해당 시즌 신규 합류 여부)로 구성하였다. 모든 선수 지표는 시즌·선수 간 비교 가능성을 확보하기 위해 90분 기준으로 정규화하였다.

팀·시즌 수준 설명변수는 Transfermarkt의 이적시장 자료와 팀 성과 이력을 바탕으로 구성하였다. 금액 관련 항목(지출, 수입, 순지출)은 로그 변환 후 해당 시즌 분포 기준으로 표준화하여 사용하였고, 선수단 구조와 변동을 나타내기 위해 (1) 영입 선수 평균 연령, (2) 방출 선수 평균 연령, (3) 전년 대비 잔류 비율(스쿼드 지속성), (4) 해당 시즌 신규 합류 비중, (5) 전년도 스쿼드 중 방출 비중을 포함하였다. 팀 성과 이력으로는 (6) 직전 시즌 승점, (7) 직전 대비 승점 변화량(전전 시즌과의 차), (8) 최근 3시즌 승점의 산술평균, (9) 최근 최대 5시즌 승점의 지수이동평균 (α=0.5)을 사용해 장·단기 추세를 함께 반영하였다.

정답 레이블은 각 시즌의 최종 리그 테이블(팀별 승점과 최종 순위)로부터 추출하였다. 전처리 측면에서 연속형 변수는 학습 세트 기준으로 표준화하였으며, 포지션과 신규 합류 여부는 범주형·이진 특성으로 반영하였다. 이러한 구성은 선수 단위의 미시적 성과, 이적시장에 따른 자원 배분, 스쿼드 구조의 연속성·변동성, 그리고 팀의 최근 성과 추세를 균형 있게 통합하도록 설계되었다.

2. 모델 구성

모델은 선수 인코더 → 시즌 내 집합 풀링 → 시간 요약 → 팀 스칼라 결합 → 회귀 헤드로 연결된다. 각 시즌의 선수 입력은 공유 MLP 인코더로 임베딩하고, 이를 순열·로스터 크기 변화에 불변인 집합 풀링으로 요약한다. 풀링은 출전 시간 기반 가중 평균과 출전 시간 상위 선수들의 평균(Top-K prior)을 결합해 시즌 임베딩을 구성하며, K는 팀별 선수단 규모에 비례하도록 설정한다. 이렇게 얻은 시즌 임베딩을 시간 요약으로 압축하고 팀·시즌 스칼라와 결합한 뒤, 회귀 헤드로 최종 예측을 산출한다 [2].

$$s_{c,t} = \left[\sum_{p \in P_{c,t}} \frac{m_{p,t}}{\sum_{q \in P_{c,t}} m_{q,t}} z_{p,t} ; \frac{1}{\left| P_{c,t}^{(20)} \right|} \sum_{p \in P_{c,t}^{(20)}} z_{p,t} \right] \quad (1)$$

여기서 $z_{p,t}$는 선수 인코더의 출력이며, $[a;b]$는 벡터 연결(concatenation)을 의미한다. 앞의 첫 항은 스쿼드 전반의 폭(뎁스)을, 두 번째 항은 핵심 선수군의 영향력을 반영한다.

시간 축에서는 대상 시즌 t 직전의 시즌 임베딩들을 단순 통계로 분해하여 장기 추세, 최신 상태, 단기 변동 신호를 함께 보존한다. 이를 위해 지수이

동평균(계수 α), 최근 시즌 값, 직전 대비 변화량을 사용하고, 세 표현을 연결해 시간 요약 임베딩을 만든다.

$$h_{c,t} = [\mathrm{EMA}_{\alpha}(\{s_{c,\tau}\}_{\tau \le t-1}); s_{c,t-1}; s_{c,t-1} - s_{c,t-2}] \quad (2)$$

이 시간 요약 임베딩에 팀·시즌 스칼라 $S_{c,t}$를 연결해 팀 차원의 구조적 정보를 통합한 후 회귀 헤드가 대상 시즌의 승점을 추정한다. 동일 시즌 내 모든 구단의 예측 승점을 내림차순 정렬해 예측 순위를 도출한다.

$$\widehat{y_{c,t}} = g_{\phi}([h_{c,t}; S_{c,t}]), \mathrm{rank}_{\cdot,t} = \mathrm{argsort}_{\downarrow}(\{\widehat{y}_{\cdot,t}\}) \quad (3)$$

위와 같은 연쇄 구성은 선수 순서 및 로스터 크기 변화에 대한 집합 불변성을 확보하면서, 핵심 선수군과 스쿼드 뎁스를 분리 반영하고, 장단기 추세와 팀 차원의 맥락을 일관된 표현으로 융합하도록 설계되었다.

3. 학습 단계

본 모델은 팀 - 시즌 단위 입력을 사용하며, 대상 시즌 t의 예측을 위해 직전 여러 시즌의 선수 텐서와 팀·시즌 스칼라 S를 활용한다. 연속형 피처는 학습 세트 통계로 표준화하고 동일 파라미터를 검증·추론에 일관 적용한다. 선수 피처는 공유 인코더로 임베딩한 뒤, 시즌 내에서는 가중 평균과 Top-K prior를 결합해 하나의 시즌 임베딩으로 집약한다. 이어 EMA·Last·Delta로 시간 정보를 요약하고, 이를 S와 연결하여 얕은 회귀 헤드로 시즌 승점을 예측한다. 학습은 표준 손실·최적화 기법을 사용하고 워크-포워드 검증, 조기 종료, 다중 시드 앙상블로 안정성과 일반화를 확보한다. 최종 순위는 예측 승점을 기준으로 정렬해 산출한다.

4. 성능 검증

본 논문에서는 모델의 성능을 검증하기 위해 대상 시즌의 직전 3개 시즌을 학습 데이터로 사용하고, 해당 시즌 프리미어리그에 참가한 각 팀의 승점 및 순위를 예측한 뒤 내림차순 정렬로 예측 순위를 산출하였다. 이후 시즌별 예측 순위와 실제 최종 순위를 비교했으며, 평가 지표로는 평균 제곱근 오차(Root Mean Squared Error, RMSE)를 사용하였다. 예를 들어 2018/2019 시즌의 경우 2015/2016-2017/2018 시즌까지의 데이터를 training set으로 삼았으며 예측한 순위와 실제 순위 사이의 오차로 계산했을 때 RMSE 값이 약 2.864로 나왔다. 해당 시즌의 팀별 예측 결과와 실제 순위 비교는 표 2에 간략히 제시하였다. 동일한 프로토콜을 나머지 시즌에도 적용하였으며, 결과는 표 1에 정리하였다.

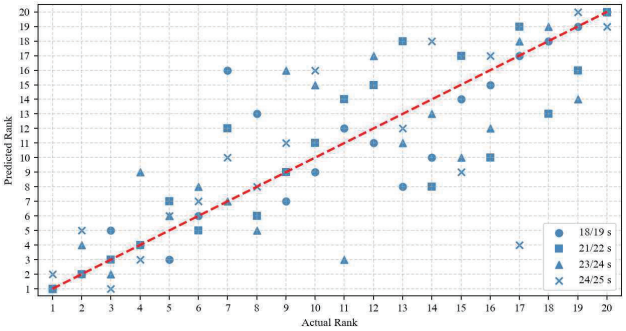

그림 1. 시즌 실제 순위(x축)와 예측 순위(y축) 비교 산점도

2019/2020 시즌은 코로나19 팬데믹으로 인한 시즌 중단·재개, 안전 프로토콜에 따른 가용 전력 변동 등 비정상적 운영 환경이 겹치며 오차가 확대되었다. 또한 2022/2023 시즌의 경우 Chelsea FC와 Newcastle United FC에서 상대적으로 큰 오차가 관측되었는데, 두 구단 모두 해당 기간 전후로 구단주 변경 및 이적정책 급변과 같은 구조적 변화가 있었고, 이는 과거 추세 기반의 팀·시즌 스칼라 및 선수 집합 신호만으로는 충분히 포착하기 어려웠다. 이러한 예외적 외생 변동을 제외하면 대부분의 시즌에서 예측 순위는 실제 순위와 비교적 잘 부합하였으며, 그림 1의 산점도에서도 예측 값이 대체로 y = x 부근에 분포함을 확인할 수 있다. 이는 선수·이적·팀 수준 정보를 결합한 본 집합 기반 접근이 시즌 최종 테이블의 상대적 위치를 실용적인 정확도로 포착함을 보여준다.

시즌	RMSE
2018/2019	2.864
2019/2020	5.254
2020/2021	4.062
2021/2022	3.098
2022/2023	5.991
2023/2024	3.741
2024/2025	4.087
평균	4.157

표 1. 시즌별 예측 성능 (RMSE)

Team	예측 순위	실제 순위	오차
Arsenal	3	5	-2
Bournemouth	10	14	-4
Burnley	14	15	-1
Chelsea	5	3	2
CrystalPalace	11	12	-1
Everton	13	8	5
Leicester	7	9	-2
Liverpool	2	2	0
ManchesterCity	1	1	0
ManchesterUtd	6	6	0
Newcastle	8	13	-5
Southampton	15	16	-1
Tottenham	4	4	0
Watford	12	11	1
WestHam	9	10	-1
Brighton	17	17	0
Cardiff	18	18	0
Fulham	19	19	0
Huddersfield	20	20	0
Wolverhampton	16	7	9

표 2. 18/19시즌 순위 예측 결과

Ⅲ. 결론

본 연구는 선수 단위 지표와 이적·팀 수준 정보를 통합해 시즌 최종 테이블(승점·순위)을 예측하는 일반 프레임워크를 제시하고, 프리미어리그를 사례로 타당성을 검증하였다. 모델은 공유 MLP로 선수 표현을 학습한 뒤, 출전 시간 가중 평균과 Top-20(출전 시간 기준) 집합 요약을 결합해 시즌 임베딩을 구성하고, 이를 EMA - Last - Delta 시간 요약과 팀·시즌 스칼라와 함께 사용해 승점을 회귀한 후 순위를 산출한다. 워크-포워드 평가에서 7개 시즌의 순위 RMSE는 평균 4.157, 최저 2.864로 나타나, 복잡한 시계열 구조 없이도 실용적 정확도를 확보함을 확인하였다. 오차가 크게 나타난 일부 시즌(예: 팬데믹, 구단 지배구조·이적정책 급변)은 외생 충격의 영향이 컸으나, 전반적으로 예측 순위는 실제 순위와 양호하게 부합했다. 제안 접근은 핵심 선수군과 스쿼드 뎁스를 분리 반영해 해석 가능성을 제공하며, 구현이 간결해 재현성과 확장성이 높다. 본 프레임워크는 특정 리그에 한정되지 않고 팀 중심 스포츠 리그 전반으로의 적용이 가능하며, 향후 사건 기반 특성(부상·징계·일정 혼잡도 등) 통합과 예측 불확실성 정량화를 통해 성능 및 실무 활용성을 더욱 향상시킬 수 있다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2025-RS-2021-II211835).

참 고 문 헌

[1] Lee, Sangmin, Hyeontae Joo, and Hwangnam Kim. "TiDoSGAN: Enhanced Generative Adversarial Network for NFC Restoration." IEEE Internet of Things Journal (2025).

[2] Bueno, Christian, and Alan Hylton. "On the representation power of set pooling networks." Advances in Neural Information Processing Systems 34 (2021): 17170-17182.