Journal of the Korean Data Analysis Society (January 2024), 26(1), 0000-0000.
https://doi.org/10.37727/jkdas.2024.26.1.0000

선수별 기여도 바탕 EPL 경기 결과 예측 모델

김지헌[1], 전현성[2], 김현규[3], 임영주[4], 정의찬[5], 박준범[6]

요 약

본 연구는 잉글리시 프리미어리그(EPL)의 경기 결과를 예측하기 위해 선수별 기여도 기반의 인공신경망 모델을 개발하였다. 기존의 팀 통계 중심 모델들이 선수 개개인의 역할을 간과하는 한계를 보완하고자, 본 연구는 공격 기여도(AC), 수비 기여도(DC), 경기 기여도(MI)를 포함하는 다차원적 지표를 설계하여 이를 바탕으로 경기 예측 모델의 입력으로 활용하였다. 선수 기여도는 팀 대비 상대적 비율로 정량화되었으며, 이를 기반으로 팀별 종합 기여도 점수를 산출하여 홈팀과 원정팀의 차이를 모델에 입력하였다. 예측 모델은 다층 퍼셉트론(MLP) 구조의 인공신경망으로 설계되었으며, 약 1,000건 이상의 EPL 경기를 학습 데이터로 사용하였다. 모델 성능은 정확도, 정밀도, 재현율 등 주요 지표를 통해 평가하였으며, 기존 단순 통계 기반 모델에 비해 높은 예측력을 보였다. 본 연구는 선수의 개별 역량을 반영한 경기 예측 가능성을 제시함으로써, 스포츠 데이터 분석 분야에서 보다 세분화된 접근 방법을 제안한다.

주요용어 : 선수 기여도, 분류, 다층 퍼셉트론

1. 서론

잉글랜드 프리미어리그(EPL)는 전 세계적으로 가장 인기 있는 축구 리그 중 하나로, 매 시즌 수십억 명의 팬들이 경기를 시청한다. 특히, 글로벌 방송 중계권 판매와 스폰서십 계약을 통한 수익 증가는 EPL이 세계 최고의 축구 리그로 자리매김하는 데 기여하고 있다. 이처럼 대중의 높은 관심 속에 방대한 경기 데이터가 생성되면서, 이를 활용한 다양한 연구 또한 활발히 진행되고 있다. 기존 EPL 경기 결과 예측 연구들은 랜덤 포레스트(Random Forest)나 로지스틱 회귀(Logistic Regression) 등의 전통적 머신러닝 기법을 활용해 팀 단위 지표나 일부 핵심 선수 데이터를 기반으로 이루어졌다. 그러나 이들 연구는 예측 정확도가 48-52% 수준에 그치거나, 단순히 변수 간 통계적 유의성 검증에 그치는 한계가 있다. 이에 본 연구에서는 최근 5개 시즌 동안의 EPL 경기 데이터를 분석하여, 개별 선수들의 종합 기여도 지표와 팀의 최근 경기력을 결합한 다중 특성 기반 머

[1]충청남도 서산시 해미면 한서1로 46, 한서대학교 항공컴퓨터학과 학부재학생. E-mail: jiheon0621@naver.com
[2]충청남도 서산시 해미면 한서1로 46, 한서대학교 항공컴퓨터학과 학부재학생. E-mail: pemjmr@naver.com
[3]충청남도 서산시 해미면 한서1로 46, 한서대학교 항공컴퓨터학과 학부재학생. E-mail: hgkim8721@naver.com
[4]충청남도 서산시 해미면 한서1로 46, 한서대학교 항공컴퓨터학과 학부재학생. E-mail: yyji81@naver.com
[5]충청남도 서산시 해미면 한서1로 46, 한서대학교 항공컴퓨터학과 학부재학생. E-mail: dmlcks74100@naver.com
[6](교신저자)충청남도 서산시 해미면 한서1로 46, 한서대학교 AI모빌리티학과 조교수. E-mail: jbpark@hanseo.ac.kr
[접수 2025년 6월 10일; 수정 2024년 1월 00일; 게재 확정 2024년 1월 00일]

신러닝 모델을 제안한다. 선수별 세부 스탯과 평점 데이터를 기반으로 한 기여도 산출, 포지션별 중요도 반영, 최근 폼(Form) 지표 통합, 그리고 상대전적 분석을 통해 기존 모델 대비 향상된 예측 성능을 목표로 한다.

2. 연구 배경

EPL은 연간 60억 파운드 이상의 수익을 창출하는 세계 최대 축구 리그로, 전 세계 200개국 이상에서 중계되고 있다. 이러한 글로벌 인기에 힘입어 스포츠 베팅 시장 규모는 연간 1조 원 이상으로 성장하였으며, 방대한 경기 데이터가 생성되면서 이를 활용한 상업적, 학문적 연구의 가능성이 급격히 증가하고 있다. 경기 데이터 기반 분석의 확산과 필요성은 현대 축구의 특징을 잘 보여준다. Expected Goals(xG), Pass Completion Rate, 압박 성공률 등 고도화된 통계 지표가 도입되면서 경기 분석의 정교함이 크게 향상되었다. 이러한 데이터는 구단의 선수 영입 및 전술 분석, 팬들의 판타지 축구 게임, 베팅 업계의 정확한 배당률 산정 등 다양한 분야에서 활용되고 있다. 본 연구에서는 선수별 종합 기여도 지표를 개발하여 예측 정확도를 높이는 것을 목표로 한다.

2.1. 데이터 수집 및 전처리

본 연구에서 사용된 프리미어리그(EPL) 경기 데이터는 FootyStats(https://footystats.org)로부터 수집되었으며, 비상업적 학술 연구 목적으로의 활용에 대해 사전 이메일을 통해 정식 허가를 받았다. 데이터는 2020년부터 2025년까지의 최근 5개 시즌을 포함하고 있으며, 각 시즌은 38라운드로 구성되어 총 약 1,900경기(시즌당 10경기 기준)에 해당한다. 수집된 데이터는 다양한 수준의 세부 정보를 포함하고 있다. 구체적으로, 각 경기의 선발 라인업 및 선수 교체 내역, 선수별 주요 경기 지표 16개 항목(예: 골, 어시스트, 키패스, 태클, 세이브 등), 각 경기의 최종 결과(홈팀 승, 무승부, 원정팀 승), 그리고 팀별 최근 경기 결과 및 득실점 기록이 포함된다. 이러한 데이터 구성은 모델이 경기 맥락과 선수 및 팀의 성과를 종합적으로 반영할 수 있도록 하는 데 기여한다. 데이터는 모델 학습 및 성능 검증을 위해 시간 순서를 기준으로 학습 세트와 테스트 세트로 분할되었으며, 분할 비율은 80:20으로 설정하였다. 수집 및 전처리 과정은 Python 기반의 FastAPI를 활용하여 자동화된 파이프라인을 통해 수행되었다.

2.2. 선수별 기여도 산출

표1은 각 선수의 종합 기여도는 포지션별 특화된 가중치를 나타낸 것이다. 경기 결과 예측의 정확도를 높이기 위해서는 개별 선수 수준의 데이터를 종합하여 팀 단위의 전력과 특성을 정량화하는 것이 중요하다. 본 연구에서는 선수별 기여도 정보를 바탕으로 다양한 팀 수준 특성을 구성하였으며, 이는 경기 전 예측 모델의 입력 변수로 사용되었다. 구체적으로 다음과 같은 지표들을 생성하였다. 먼저, 평균 기여도는 선발 출전한 11명의 선수에 대한 기여도(공격, 수비, 패스 등 종합 지표)의 평균값으로 계산하였다. 여기에 교체 명단에 포함된 선수들의 기여도는 경기 영향력이 상대적으로 낮음을 고려하여 0.2의 가중치를 부여한 후 함께 평균에 반영하였다. 이를 통해 선발 라

인업 중심의 팀 전력을 정량적으로 평가할 수 있도록 하였다.

Table 1. Estimating individual player contributions

Position	Contribution Type	Formula
Forward	Attacking contribution (AC)	0.45×Goals/90 + 0.25×xG/90 + 0.15×Assists/90 + 0.05×Shots/90 + 0.10×Finishing efficiency
Midfielder	Attacking contribution (AC)	0.3×Goals/90 + 0.3×Assists/90 + 0.4×Key passes/90
Midfielder	Defensive contribution (DC)	0.6×Tackles/90 + 0.4×Interceptions/90
Defender	Defensive contribution (DC)	0.3×Tackles/90 + 0.3×Interceptions/90 + 0.25×Clearances/90 + 0.15×Aerial duel win rate/90
Goalkeeper	Defensive contribution (DC)	0.4×Saves/90 + 0.4×Clean sheets/90 − 0.2×Goals conceded/90
All positions	Match impact (MI)	Team win rate when the player is on the pitch
Final score	Total contribution	0.4×AC + 0.4×DC + 0.2×MI (Normalized to 0 – 100)

2.4. 실험 설정 및 성능 평가

본 연구에서는 예측 모델의 학습과 평가를 위해 체계적인 전처리 및 학습 절차를 수행하였다. 전체 프로세스는 (1) 시간 순서 기반의 80:20 데이터 분할, (2) 결측치의 0 대체, (3) Grid Search를 활용한 하이퍼파라미터 튜닝, (4) 5-fold 교차 검증의 네 단계로 구성된다. 시간 기반 분할은 실제 시나리오를 반영하여 과거 데이터를 학습에, 최신 데이터를 평가에 활용하였으며, 결측값은 정보 부재 상황을 고려해 보수적으로 처리하였다. 최적의 모델 구성을 위해 Grid Search를 적용하였고, 교차 검증을 통해 모델의 일반화 성능과 안정성을 확보하였다. 이러한 절차를 통해 본 연구는 과적합을 방지하고 현실적인 예측 성능을 달성하고자 하였다.

Table 2. Overall performance comparison

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.481	0.468	0.451	0.448
Random Forest (Basic)	0.481	0.495	0.481	0.479

Random Forest (Advanced)	0.540	0.532	0.540	0.528
Proposed Model	0.583	0.574	0.583	0.571

표 2는 본 연구에서 제안한 모델과 기존 기법(로지스틱 회귀, 기본/고급 랜덤 포레스트) 간의 전반적인 성능을 Accuracy, Precision, Recall, F1-score 네 가지 지표를 기준으로 비교한 결과를 나타낸다. 기존 연구에서 주로 활용된 로지스틱 회귀(Logistic Regression)와 기본(Random Forest - Basic) 모델은 정확도(Accuracy) 기준 각각 0.481로 유사한 수준을 보였으며, 정밀도(Precision)와 재현율(Recall) 역시 상대적으로 낮은 값을 기록하였다(각각 0.468 및 0.451, 또는 0.495 및 0.481). 고급(Random Forest - Advanced) 모델의 경우, 하이퍼파라미터 튜닝과 추가적인 피처 엔지니어링이 반영되면서 Accuracy 0.540, F1-score 0.528로 성능이 향상되었음을 확인할 수 있다. 그러나 여전히 실제 적용을 고려하기에는 예측력이 부족한 수준에 머무르고 있다.

반면, 본 연구에서 제안한 모델(Proposed Model)은 Accuracy 0.583, Precision 0.574, Recall 0.583, F1-score 0.571로, 모든 평가 지표에서 기존 모델들을 상회하는 성능을 보였다. 이는 선수별 종합 기여도, 포지션별 가중치, 팀의 최근 폼(Form) 및 상대 전적 정보를 반영한 다중 특성 기반 접근 방식이 경기 결과 예측의 정확도를 실질적으로 개선하는 데 기여했음을 시사한다. F1-score에서의 향상은 단순히 정확도만이 아니라 클래스 간 균형 있는 예측 성능을 확보했음을 의미하며, 실제 스포츠 경기 예측과 같이 클래스 불균형(class imbalance)이 존재할 수 있는 상황에서 더욱 중요한 지표로 간주된다. 종합적으로 볼 때, 제안된 모델은 기존 기법 대비 높은 신뢰성과 일반화 성능을 확보하며, 실제 응용 가능성이 높은 예측 모델임을 실험적으로 입증하였다.

3. 연구 결과

본 연구에서는 EPL 최근 5개 시즌의 경기 데이터를 기반으로, 선수별 종합 기여도, 포지션별 가중치, 팀의 최근 폼 및 상대 전적 등 다양한 특성을 통합한 머신러닝 기반 예측 모델을 제안하였다. 실험 결과, 제안된 모델은 기존 로지스틱 회귀 및 랜덤 포레스트 모델 대비 모든 평가 지표(Accuracy, Precision, Recall, F1-score)에서 향상된 성능을 보였으며, 특히 정확도는 0.583, F1-score는 0.571로 가장 높은 값을 기록하였다. 이는 팀 단위가 아닌 개별 선수의 특성과 경기 맥락을 종합적으로 반영한 접근이 경기 결과 예측의 정확도를 높이는 데 효과적임을 시사한다. 향후 연구에서는 다음과 같은 방향으로 확장을 고려할 수 있다. 첫째, 경기 중 실시간 데이터(예: 선수 움직임, 패스 맵, 부상 정보 등)를 활용한 동적 예측 모델 개발이 필요하다. 둘째, 딥러닝 기반의 시계열 모델(LSTM, Transformer 등)을 적용하여 경기 흐름을 보다 정밀하게 반영하는 방법도 고려할 수 있다. 셋째, 타 리그나 국가대표 경기 등 다양한 데이터셋에 본 모델을 적용하여 일반화 가능성을 검증하는 것도 유의미한 연구가 될 것이다.

Reference

Yoon, Y. I., Jeong, H. Y. (2024). A Comparison of Uncertainty Quantification of Deep Learning models for Time Series, Journal of the Korean Data Analysis Society, 26(1), 163-174. (in Korean)

Oh, J. M., Shin, H. S., Shin, Y. S., Jeong, H. C. (2017). Forecasting the Particulate Matter in Seoul using a Univariate Time Series Approach, Journal of the Korean Data Analysis Society, Vol. 19, No. 5 (B), pp. 2457-2468. (in Korean)

Park, M. S. (2019). Regional Association of the Particulate Matters, Journal of the Korean Data Analysis Society, Vol. 21, No. 3, pp. 1169-1181. (in Korean)

Kwak, S. J., Cho, J. I., Choi, E. C. (2022). Fine Dust is Coming Again: The Effect of Air Pollution on Health Using Seasonal Weather Patterns, Journal of the Korean Data Analysis Society, 24(5), 1625-1637.

Lee, D. H., Yang, H. J. (2023). Estimation and Prediction Model for Functional Time Series Data and Application, Journal of the Korean Data Analysis Society, 25(5), 1713-1723. (in Korean)

EPL match result prediction model based on player contribution

Jiheon Kim[1], *Hyunseong Jeon*[2], *HyeonGue Kim*[3], *Youngju Im*[4], *EuiChan Jeong*[5], *JunBum Park*[6]

Abstract

This study proposes a neural network model for predicting match outcomes in the English Premier League (EPL) by incorporating player-specific contribution metrics. Unlike traditional models that rely heavily on team-level statistics, our approach quantifies individual player performance through three core dimensions: Attacking Contribution (AC), Defensive Contribution (DC), and Match Involvement (MI). Each metric is normalized by the team's overall performance, enabling a relative assessment of each player's impact. These aggregated team-level contributions are then used as input features for a multi-layer perceptron (MLP) architecture. The model is trained on over 1,000 EPL matches and evaluated using standard metrics such as accuracy, precision, and recall. Experimental results demonstrate that the proposed model outperforms conventional baseline models, highlighting the predictive power of incorporating fine-grained player-level data. This research underscores the significance of individualized performance metrics in advancing sports analytics and outcome forecasting.

Keywords : Player Matrix, Classification, Multi-Layer Perceptron

[1]Undergraduate Student, Department of Aerospace Computer Engineering, Hanseo University, Seosan, 31962 South Korea.
E-mail: jiheon0621@naver.com
[2]Undergraduate Student, Department of Aerospace Computer Engineering, Hanseo University, Seosan, 31962 South Korea.
E-mail: pemjmr@naver.com
[3]Undergraduate Student, Department of Aerospace Computer Engineering, Hanseo University, Seosan, 31962 South Korea.
E-mail: hgkim8721@naver.com
[4]Undergraduate Student, Department of Aerospace Computer Engineering, Hanseo University, Seosan, 31962 South Korea.
E-mail: yyji81@naver.com
[5]Undergraduate Student, Department of Aerospace Computer Engineering, Hanseo University, Seosan, 31962 South Korea.
E-mail: dmlcks74100@naver.com
[6](Corresponding Author) Assistant Professor, Department of AI Mobility, Hanseo University, Seosan, 31962 South Korea.
E-mail: jbpark@hanseo.ac.kr
[Received 00 January 2024; Revised 00 January 2024; Accepted 00 January 2024]