영국 프리미어 리그 경기 데이터를 이용한 CatBoost 기반의 경기 결과 예측

이용선[1], 이지민[1], 조현수[2], 남현준[2], 김보영[2], 문지훈[1]
[1] 순천향대학교 AI·빅데이터학과
[2] 아산중학교
[1] {20211487, dlwlals7359, jmoon22}@sch.ac.kr
[2] {astre6099, namnam0226, bboda98}@gmail.com

CatBoost-Based Match Outcome Prediction Using English Premier League Match Data

Yong-Sun Lee[1], Jimin Lee[1], Hyun-Su Cho[2], Hyun-Jun Nam[2],
Bo-Young Kim[2], and Jihoon Moon[1]
[1] Department of AI and Big Data, Soonchunhyang University
[2] Asan Middle School

요 약

축구는 전 세계 스포츠 시장에서 중요한 위치를 차지하며, 수많은 팬의 관심과 열정을 받아왔다. 이러한 인기와 더불어 많은 연구자와 분석가들이 경기의 승패를 예측하려는 시도를 해왔지만, 축구의 다양한 변수와 불확실성으로 인해 예측의 정확성은 항상 도전적인 부분이었다. 최근 빅데이터 및 인공지능 기술의 발전으로 스포츠에서도 기계학습을 이용하여 경기의 결과를 예측하는 연구가 보고되고 있다. 본 연구는 경기 기록 데이터를 활용하여 CatBoost (Categorical Boosting) 기반의 예측 모델을 구성하고, 이를 통해 경기 결과를 예측하는 방안을 제시한다. 또한, SHAP (SHapley Additive exPlanations)을 활용하여 각 변수의 중요도를 분석하여, 예측의 신뢰성을 높이는 방향을 모색한다.

1. 서 론

축구는 국경을 넘어 전 세계적으로 많은 인기를 보유한 스포츠로, 여러 연구자와 팬들 사이에서 경기 결과 예측에 관한 관심이 끊임없이 높아져 왔다[1]. 그러나 다른 종목들과 달리 축구는 득점이 상대적으로 적게 일어나고, 무승부의 발생 확률이 높은 특성으로 인해 결과 예측의 정확도가 떨어질 수 있다는 한계가 존재한다[2].

최근 4차 산업 혁명의 원동력으로 인해 인공지능 기술 발전이 가속화되면서, 스포츠 분야에서도 데이터 기반의 기계학습 모델이 새로운 도전과 기회를 제시하고 있다[3]. 이런 맥락에서, 본 연구는 전 세계적으로 높은 시청률과 관심을 받는 영국 프리미어 리그(EPL)의 경기 데이터를 중심으로, 기계학습 기법을 활용한 축구 경기 결과 예측 모델을 제시한다.

본 논문의 나머지는 다음과 같이 구성된다. 2장에서는 본 연구에서 활용된 데이터 셋의 구조를 소개한다. 3장은 승부 예측을 위해 사용된 기계학습 모델에 관해 설명한다. 4장에서는 해당 예측 모델의 분석 결과 및 성능 평가를 진행한다. 마지막으로 5장에서는 연구 결론 및 향후 연구 방향에 대해 논의한다.

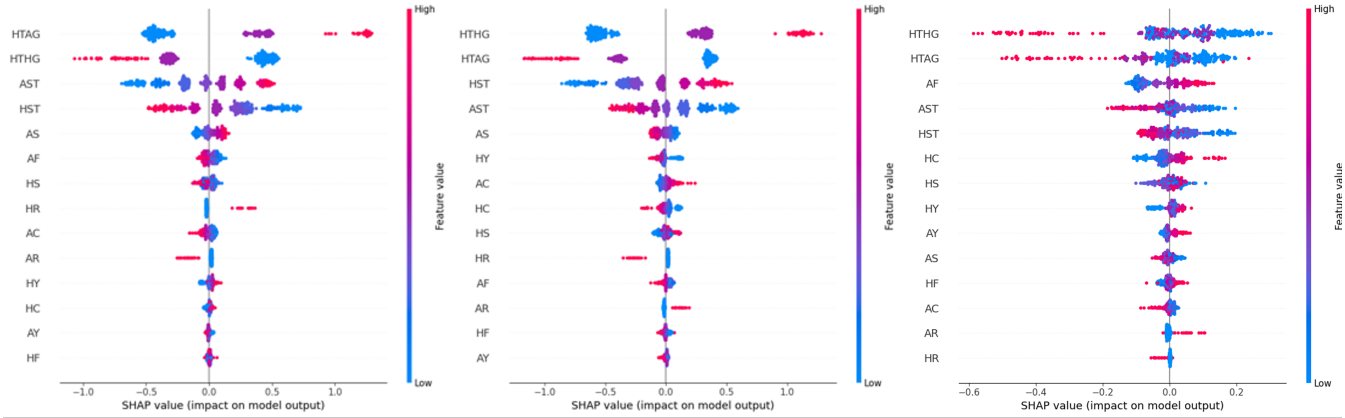
2. 데이터 셋

본 연구는 영국 프리미어리그(EPL)의 경기 결과를 예측하고자 2000-2001시즌부터 2020-2021시즌까지의 총 21개 시즌 동안의 경기 기록 데이터를 학습 데이터로 활용한다. 또한, 2021-2022시즌의 경기 기록은 모델의 성능을 검증하기 위한 평가 데이터로 사용한다.

본 연구는 여러 경기 관련 변수들을 모델의 입력으로 사용하며, 대표적인 변수로 유효슈팅의 개수, 반칙의 횟수, 코너킥의 횟수 등이 포함되어 있다. 이러한 변수들의 상세 리스트는 <표 1>에 나타내었다. 또한, 학습 데이터에는 있으나 평가 데이터에는 포함되지 않은 팀들의 경기 기록 데이터는 분석에서 제외한다. 각 팀에 대한 정보는 One-Hot Encoding 방법을 통해 총 20개의 팀 데이터를 60개의 변수로 변환하여 사용한다. 모델의 출력 변수로는 경기의 결과를 '홈팀 승리', '원정팀 승리', '무승부'의 세 가지 카테고리로 범주화하여 활용한다.

(표 1) 경기 예측 모델의 학습을 위한 독립변수 목록

분류	변수
범주형 변수	Home-team NO.1 ~ NO.20, Away-team NO.1 ~ NO.20, Full-Time Result
수치형 변수	Season, Date Time, Full-time Home-team goals, Full-time Away-team goals, Half-time Home-team goals, Half-time Away-team goals, Home-team Shooting, Away-team shooting, Home-team Shoot in target, Away-team Shoot in target, Home-team corner kick, Away-team corner kick, Home-team Foul, Away-team Foul, Home-team Yellow card, Away-team Yellow card, Home-team Red card, Away-team Red card
기타 변수	Referee name, Half-time Result



(그림 1) CatBoost 모델 예측에 관한 SHAP의 Summary Plot 시각화(왼쪽: 홈팀 승리, 가운데: 원정팀 승리, 오른쪽: 무승부)

3. 분류 모델 구성

CatBoost (Categorical Boosting)는 Gradient Boosting 기반의 기계학습 기법으로, 특히 범주형 데이터의 처리에 있어 뛰어난 성능을 발휘한다. 이 기법은 과적합을 효과적으로 방지하는 메커니즘 및 GPU 지원 등과 같은 유용한 기능들로 구성되어 있으므로 복잡한 데이터 패턴 학습에 적합하다[6]. SHAP (SHapley Additive exPlanations)은 기계학습 모델 예측에 관한 설명력을 제공하는 도구로써, 각 특성이 예측에 어느 정도 영향을 미쳤는지 정량적으로 평가할 수 있다[7].

따라서, 본 연구는 범주형 데이터인 종속변수의 예측을 위해 CatBoost 알고리즘을 적용한다. 또한, 예측 모델에서 각 변수의 영향력을 분석하고자 SHAP을 활용하여 주요 변수들의 기여도를 파악한다.

4. 실험 결과

본 연구에서 학습한 CatBoost 모델을 통해 얻은 예측 결과는 표 2와 같다. 제안한 모델의 전반적인 정확도는 68%로 확인하였으며, 승리 팀에 관한 예측은 상대적으로 정확하였으나, 무승부에 관한 예측은 다소 정확하지 않은 경향을 확인하였다.

(표 2) 경기 예측에 관한 혼동 행렬 결과

Category	Precision	Recall	F1-score	Support
홈팀 승리	0.74	0.76	0.75	107
원정팀 승리	0.71	0.78	0.75	130
무승부	0.47	0.38	0.42	72
정확도		0.68		309

본 연구에서 SHAP 도구를 활용하여 CatBoost 모델의 예측값에 기여하는 주요 변수들을 분석하였으며, 그림 1과 같이 상세하게 나타내었다.

홈팀 승리를 예측하는 데 가장 큰 영향을 미친 변수는 홈팀의 전반전 득점이었다. 홈팀이 전반전에 많은 득점을 할수록 홈팀 승리로 예측하는 확률이 상승하였다. 반대로 원정팀 승리를 예측하는 데 주요 변수는 원정팀의 전반전 득점으로 확인하였으며, 원정팀의 득점이 많을 때 원정팀 승리로 예측하는 확률이 높아졌다. 무승부의 예측에서는 양 팀의 전반전 득점이 주요 변수로 작용하였다. 이뿐만 아니라, 각 팀의 승리 예측에서는 유효슈팅이 두 번째로 중요한 변수로 나타났지만, 무승부 예측에서는 원정팀의 파울 횟수가 두 번째로 큰 기여도를 보였다. 이를 통해, 특정 변수들이 경기 결과 예측에 있어 중요한 역할임을 확인할 수 있었다.

5. 결 론

본 연구에서는 EPL 경기 결과를 예측하고자 CatBoost 기법을 적용하고, SHAP을 통해 각 변수의 기여도를 파악하였다. CatBoost 모델은 68%라는 높은 예측 정확도를 달성하였으며, 전반전 득점 수가 종속변수의 모든 범주에 걸쳐 가장 큰 영향을 미쳤음을 확인하였다.

향후 다양한 기계학습 모델에 대한 탐색을 통해 단순한 모델 적용을 넘어서 최적화 기법을 도입하여 더욱 강건한 예측 모델을 구성하고자 한다.

사 사 문 구

본 연구는 2021년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음(2021-0-01399).

참 고 문 헌

[1] 김형원, 구명완, “최근 경기 데이터를 활용한 EPL 경기 예측”, 『한국정보과학회 학술발표논문집』 2019, pp. 748-750.

[2] 김필수, “영국 프리미어리그 경기데이터 기반 머신러닝을 활용한 경기결과예측 및 분류모형의 예측 성능 비교”, 『한국체육학회지』 제62권, 제4호, 2023, pp. 337-353.

[3] 조정환, “스포츠 빅데이터 활용과 전망”, 『한국체육측정평가학회지』 제14권, 제2호, 2012, pp. 1-11.

[4] Hancock JT, Khoshgoftaar TM, “CatBoost for big data: an interdisciplinary review”, Journal of big data. Vol. 7, No. 1, 2020. pp. 1-45.

[5] Lundberg SM et al., “From local explanations to global understanding with explainable AI for trees”, Nature Machine Intelligence. Vol. 2, No. 1, 2020. pp. 56-67.

공지사항

스마트미디어학회 조직위원회를 대표하여, 2023 종합학술대회를 보다 잘 준비하기 위한 협력에 감사드립니다. 논문 제출과 더불어 해당 논문의 분야에 대해 조사하고자 합니다.

논문이 수락된 경우, 학회 프로그램 세션에 분류될 트랙을 선택해주시기 바랍니다.

[ ] Smart Energy ICT
AMI(지능형 계량시스템), EMS(에너지관리시스템), BEMS(건물에너지관리시스템)
스마트홈, IoT, 스마트그리드, 마이크로그리드, 송변전자동화시스템, 배전자동화시스템
MDMS(계량데이터관리시스템), 전력거래시스템, EV, 분산전원, VPP(가상발전소)
ESS(에너지저장시스템), V2G(Vehicle to Grid)

[ ] Smart Information
지능형컴퓨터, 클라우드컴퓨팅, 분산 및 병렬처리시스템, 인공지능, 영상처리
컴퓨터그래픽스, 음성처리, 멀티미디어, HCI, 빅데이터, 지능정보처리, 정보보호
모바일정보통신, 사물인터넷, 자동제어, 반도체, Microwave/Wireless, Optics

[ ] Information System
정보시스템 조직과 관리, e-비즈니스, ERP, CRM, SCM, 스마트워크, 소셜네트워크
IT아웃소싱, 프로젝트관리, 스마트라이프, 스마트 물류/금융/농업/교통/헬스케어
산업융합보안, 개인정보/의료정보/금융정보/산업기술보호, 스마트그리드, AMI

[ ] Contents & Services
인터렉티브콘텐츠, UX/UI 디자인, 서비스디자인, 디자인기반이론, 만화/애니메이션,
VR, AR, MR, 게임, 건축디자인, 문화예술콘텐츠, 관광콘텐츠, 인문사회융합콘텐츠,
Police & Law 콘텐츠, 인공지능(AI) 콘텐츠, 보건콘텐츠, 교육콘텐츠, 식품콘텐츠,
예술 콘텐츠 평론

[ ] Smart Media
미디어융합, 홀로그램, 디지털사이니지, 스토리텔링, 방송미디어, 인공지능 미디어 서비스.
미디어 정책, 메타버스, 미디어파사드

[√] Big Data
빅데이터 기술, 빅데이터 분석, 빅데이터 네트워크 및 구현, 빅데이터 정보보안,
빅데이터 비즈니스 모델, 빅데이터 컨설팅, 빅데이터 교육