JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 3, pp. 719-732, Mar. 2024

Check for updates

엘로 평점 시스템과 머신러닝 알고리즘을 적용한 선행적 한국프로축구 경기 결과 예측 및 분류모형 성능평가

김 필 수[1‡] · 이 상 현[2‡] · 서 재 현[3*]
[1‡] 한국스포츠경영전략연구원 원장 [2‡] 한국스포츠경영전략연구원 부원장 [3*]한양대학교 석사과정

Evaluating the Performance of Predictive Models Using the Elo Rating System and Machine Learning Algorithms: Empirical Evidence of the Korean Professional Football League

Philsoo Kim[1‡] · Sang Hyun Lee[2‡] · Jae Hyun Seo[3*]
[1‡] Director, Korea Sport Management Research Institute, Seoul 06543, Korea
[2‡] Deputy Director, Korea Sport Management Research Institute, Seoul 06543, Korea
[3*]Master's Course, Department of Industrial Data Engineering, Hanyang University, Seoul 04763, Korea

[요 약]

본 연구는 경기 결과의 예측을 위해 엘로 평점 시스템을 개념화하는 한편, 머신러닝 알고리즘의 적용을 통한 실증분석을 수행하기 위해 수행되었다. 이를 위해 한국프로축구 홈페이지에 구축된 K리그 2020~2023시즌의 790경기 관련 자료를 수집하여 전처리하고, 엘로 평점 시스템 관련 변수를 포함한 총 271개의 변수를 생성하였다. 이후, 생성된 변수를 대상으로 scikit-learn 라이브러리에서 제공하는 Random Forest Classifier의 Feature Importance 기능을 활용하여 경기 결과 예측에 기여도를 높이는 변수별 설명력을 측정하고, 엘로 평점 시스템 관련 변수 5가지를 포함한 총 120개의 독립변수 선별을 거쳤다. 마지막으로, Naive Bayes, Logistic Regression, Light GBM, Elastic Net, Decision Tree의 5가지 머신러닝 알고리즘 중 한국프로축구 경기 결과 예측 정확도 0.48로 가장 우수한 성능을 보인 Decision Tree 기반 예측 모델에 관해 본 연구에서 제안하는 경기 결과 예측 의사결정론을 도입함으로써 최종 예측 정확도 0.51로 한국프로축구 경기 결과를 선행적으로 예측하였다.

[Abstract]

This study was conducted to theoretically conceptualize the application of the Elo rating system for match prediction using machine learning algorithms and empirically test its role using machine learning algorithms with K-League match data for 790 games in the 2020-2023 seasons. Python (3.10.9) was used to collect and preprocess match-related data from the K-League website. Then, the importance of each variable to the prediction of game results was measured using the Feature Importance function from the Random Forest Classifier in the scikit-learn library for the generated variables. A total of 120 independent variables, including five Elo rating system related variables, were selected. Finally, among five machine learning algorithms (naive Bayes, logistic regression, light GBM, elastic net, and decision tree), the decision tree based prediction model scored the highest prediction accuracy of 0.48. This was then reinforced by our own prediction making process, yielding a prediction accuracy of 0.51 in the K-league prediction task.

색인어 : 경기 결과 예측, 머신러닝, 엘로 평점 시스템, 인공지능, 한국프로축구

Keyword : Game Result Prediction, Machine Learning, Elo Rating System, Artificial Intelligence, K-league

http://dx.doi.org/10.9728/dcs.2024.25.3.719

Received 31 January 2024; **Revised** 27 February 2024
Accepted 19 March 2024

‡ These authors contributed equally to this work
***Corresponding Author; Jae Hyun Seo**
Tel: ※ 개인정보 표시제한
E-mail: shscottlee@naver.com

Ⅰ. 서 론

4차산업혁명이 본격화되며 인공지능과 빅데이터를 통한 기술적 집적과 기존산업과의 융합의 실현으로 기존과는 다른 사고방식과 정보혁명으로 인해 새로운 의사결정과 질서가 창출되고 있다[1]. 이러한 4차산업혁명의 현상은 스포츠 산업에서도 계속해서 나타나며 변화를 주도하고 있다. 인공지능과 빅데이터의 스포츠 산업적 결합은 스포츠 애널리틱스(sports analytics)의 이론적 발전과 실증의 집적으로 4차산업혁명을 견인하고 있다.

스포츠 애널리틱스란 스포츠 관련 빅데이터를 활용하여 스포츠 산업의 다양한 이해관계자와 조직이 최선의 의사결정을 할 수 있도록 도움을 주는 과정을 의미한다[2],[3]. 스포츠 애널리틱스의 개념적 정의에는 인공지능이 포함되지는 않지만, 실시간 의사결정에 도움이 되는 데이터의 정제나 상황별 실시간 예측 시나리오의 반영을 위한 인공지능 활용 폭의 증가가 중요한 요인으로 고려된다[4],[5].

스포츠 애널리틱스는 큰 틀에서 경기와 직접 관련된 데이터를 활용한 분석(on-field analytics)과 경기 외부의 데이터를 활용한 분석(off-field analytics)으로 나눌 수 있다. 경기데이터를 직접 활용한 데이터 분석은 선수와 팀의 경기력, 전술 및 전략 분석, 플레이 스타일, 훈련 등과 같이 경기장 내에서 생성되는 데이터를 바탕으로 이루어진 분석을 의미하며, 경기 외부의 데이터를 활용한 분석은 티켓판매나 광고와 같이 경기 관련 데이터와는 직접 상관이 없는 방식의 데이터로 스포츠 구단이나 조직의 수익성에 영향을 미칠 수 있는 데이터 분석을 의미한다[2],[3].

본 연구에서 체계적으로 정리한 광의의 정의를 반추하면, 스포츠 애널리틱스가 경기 외적인 데이터 분석을 포함하지만, 일반적으로는 경기데이터를 활용한 데이터 분석을 중심축으로 이론적 발전이 가속화되며 실증의 축적이 이루어진다고 볼 수 있다[2]. 본 연구 역시 경기데이터 분석에 초점을 맞춰 논의를 진행하고자 한다. 특히, 선행연구 고찰의 차원에서 인공지능의 적용이 스포츠 산업과 현장에서 본격화되기 시작한 2010년대 중반 이후, 인공지능과 빅데이터를 활용하여 경기 결과 예측(game-prediction, sports prediction)을 진행한 연구가 급격히 증가하기 시작했다[6]-[8].

인공지능과 빅데이터를 활용한 경기 예측 연구의 중요성은 전술한 4차산업혁명의 개화와 기술적 발전 요인과 관련되어 있다. 최근 전 세계적으로 급격하게 확대되어 성장하고 있는 스포츠 베팅시장 역시 인공지능과 빅데이터를 본격적으로 활용할 수 있는 스포츠 관련 데이터의 축적, 알고리즘과 하드웨어 및 소프트웨어의 발달, 경기 결과를 예측하고 싶어 하는 인간의 본연적 욕망, 자본의 중요성이 더욱 강조되고 있는 풍조 등이 결합하여 나타난 현상이라고 볼 수 있다.

이에 따라 경기 결과 예측에 관한 연구가 축적되기 시작하면서 프로스포츠는 물론 다양한 종목에서 스포츠 애널리틱스가 적용되고 있으며, 이에 대한 데이터의 전처리, 분석에 포함되어야 하는 변수 및 딥러닝과 머신러닝 알고리즘 등에 있어서는 학술적으로 일정 수준의 합의가 이루어지고 있다고 보인다. 하지만, 스포츠 경기 결과를 인공지능과 빅데이터를 적용하여 '예측'한다는 관점에서 볼 때, 스포츠 애널리틱스를 이론적으로 적용한 경기 결과 예측과 관련하여 중요하게 고려할 연구의 맥락은 다음과 같다.

첫째, 예측(prediction, forecast)이라는 개념은 사건이 발생하기 전 결과를 추정하거나 설명한다고 볼 수 있으며, 과학적 체계성을 바탕으로 데이터 분석 과정 전반에서 과거에 벌어진 사건의 패턴을 활용한다. 하지만, 현재까지 스포츠 분야에서 수행된 대부분의 예측과 관련된 선행연구에서는, 예측을 위해 사용된 변수들과 예측을 위한 결과 변수가 동시에 활용됨으로써 일반적으로 사용되는 예측이라는 단어와 일치하지 않는다는 한계를 지닌다[9],[10]. 이는 승패가 결정되는 상당수의 스포츠 종목에서 선수 개인의 경기력을 측정하는 기록으로만 구성되는 것이 아니라 상대와의 정합성과 팀 내외의 다양한 요인의 역동에 영향을 받기 때문에 과거의 기록만을 가지고 현재의 경기 결과를 예측하기 위해서는 새로운 방식의 접근이 필요해 보인다[11].

둘째, 대다수 스포츠팬이나 산업 관계자들이 관심이 있는 스포츠 종목 중, 프로축구 경기 결과에 관한 예측 연구는 매우 드물다. 축구는 농구나 야구 종목과 비교하면 스포츠 애널리틱스 관점의 경기 결과 예측을 시도하고 실증적으로 구현하는데 많은 장애 요인이 존재한다. 프로농구 종목은 축구에 비해 많은 득점이 발생하고 선수들의 출장 시간이 초 단위로 기록될 만큼 상세한 기록의 측정과 유지를 할 수 있으므로 경기기록과 경기 결과 사이의 분석이 비교적 쉽다[12],[13]. 또한, 프로야구 경기는 대부분의 플레이 사이에 분절이 발생하고 선수들의 플레이 대부분이 상세하게 기록된다는 측면에서 경기 데이터 분석을 통한 경기 결과 예측이 더욱 정확하게 나타날 수 있다는 측면이 존재한다. 하지만, 프로축구는 플레이 사이에 분절이 다른 종목에 비해 현저하게 적을 뿐만 아니라 득점이나 득실 차가 적게 나타나며, 무승부가 많이 발생하는 특성으로 인해 인기나 시장 규모와 비교하면 경기 결과 예측에 관한 연구가 많지 않다[14].

셋째, 국내 프로축구인 K리그 경기 결과에 대한 선행적 예측은 더욱 찾기 힘든 실정이다. 한국프로축구는 역사적으로 발전된 기간과 대중적인 인기에도 불구하고, 경기에 대한 상세한 기록의 축적이 체계적으로 구축되어 있다고 보기 힘들다. 그 결과, 한국프로축구에 관해서는 극소수의 선행연구를 통해 인공지능을 적용한 알고리즘을 활용하여 경기 결과 예측을 위해 노력해 온 것으로 확인되었다[13],[14]. 국내에서 축구라는 종목의 위상과 산업적 가치를 고려한다면, 더욱 많은 연구와 분석이 프로축구 경기 데이터 분석을 통하여 이루어져야 할 것으로 보인다.

이러한 맥락에서 본 연구의 목적은 기존 연구의 한계를 극복하고 스포츠 경기가 일어나기 전에 자료를 수집, 정제, 전처리 및 분석을 진행하여 한국프로축구 리그 경기 결과를 선행

적으로 예측하고자 진행되었다. 이를 위해 2020-2023시즌(2023년 9월 24일까지)에 치러진 경기기록을 수집하고 머신러닝 알고리즘을 적용하여 선행적으로 한국프로축구 리그의 경기 결과를 예측하고자 한다.

특히, 경기에 대한 상세 기록이 한정적으로 축적된 한국프로축구 리그의 경기기록에 더해 본 연구에서는 기존의 선행연구에서 전혀 고려하지 않은 엘로 평점 시스템(Elo rating system)을 활용하여 선행적 경기 결과 예측을 위해 투입하는 게 중요하다고 판단된다. 이를 바탕으로 본 연구에서는 선도적으로 한국프로축구 경기 결과에 대해 사전적으로 예측하고자 한다. 이는 프로축구 경기 결과 예측에 대한 실증분석을 진행할 때 참고할 수 있는 하나의 이론과 방법론 측면에서의 새로운 시각을 제공하였다는 측면에서 중요한 의의를 지닐 것으로 보인다.

Ⅱ. 엘로 평점 시스템

2-1 엘로 평점 시스템과 경기결과 예측

본 연구에서는 엘로 평점 시스템에 대한 개념에 대해 체계적으로 고찰하여 경기 결과 예측에 있어 관련 변수 도입에 대한 정당성을 부여하고 이에 대한 의의를 설명하고자 한다. 엘로 평점 시스템이란 아르파드 엘뢰(Árpád Élő) 박사가 체스 경기에서 선수의 실력을 등급화하기 위해 고안된 점수 측정 방식으로 플레이어 간 상대 전적이 없는 상태에서도 상대적 실력에 대한 차이를 수치화하기 위한 지표이다.

보다 구체적으로, 엘로 평점 시스템은 시합을 펼치는 선수 간 예상 승률을 계산하여 등급을 조절하며 승패 결과나 상대적인 등급 차이에 따라 경기 결과에 입각한 등급이 조정되는 특성이 있다. 예를 들어, 비슷한 등급을 가진 선수를 상대로 승리하면 미세한 등급 조정이 이루어지지만, 등급의 차이가 상대적으로 큰 선수에게 승리할 때 등급의 조정이 크게 이루어진다. 따라서, 단순히 승률만을 표시하는 방식에 비해 선수 간 실력 차이를 반영하여 등급을 결정한다는 점에서 평점에 대한 신뢰도가 높다[15],[16].

엘로 평점 시스템은 상대적 실력에 대한 수치화로 실력을 서열화할 수 있다는 장점을 내포하기 때문에 다양한 스포츠 기구에서 순위(ranking) 산정에 활용하고 있다. 하지만, 그 신뢰성과 실무적 활용도에도 불구하고 선행연구에서는 농구, 야구, 축구 등과 같은 프로스포츠 종목의 경기 결과 예측에 실제로 많이 적용하여 사용하지는 않는 것으로 보인다[17]. 따라서, 본 연구에서는 엘로 평점 시스템을 경기 결과 예측을 위한 요인 중 하나로 사용하여 그 효과성에 대해 실증적으로 검증하고자 한다.

경기데이터를 활용하여 승패를 예측한 기존의 선행연구는 경기데이터에 포함된 다양한 지표들을 그대로 변수로 사용하거나 의미 있는 형태로 가공하고 계량하여 2차 경기지표(advanced stats)로 변환하여 분석에 사용한다. 어떤 변수를 사용할 것인지에 대한 선택은 데이터 분석 결과를 해석하는 모형화하는 자체 판단에 따라 달라질 수 있으며, 보다 의미 있는 변수를 선택할 때 경기 결과 예측의 정확성을 증가시키는 방향성으로 최적화할 수 있다[9],[14].

독립변수와 종속변수가 같은 시점에 측정되는 경우 존재하는 경기데이터를 중심으로 판단할 수 있으므로 비교적 객관적인 판단을 내릴 수 있다. 하지만, 선행적 예측의 경우 선수나 팀의 현재 기록이 다음 경기에서 결정적인 역할을 할 수 있을지에 관한 판단이 어려울 뿐 아니라, 상대성의 측면에서도 달라질 수 있다는 문제로 인해 어떤 변수를 선택하여 미래를 예측할 것인지에 대해 연구자의 판단이 개입할 여지가 크다.

이러한 측면에서 경기 결과를 예측하기 위해 엘로 평점 시스템의 적용은 승패 예측의 정확성을 개선하기 위해 객관적인 도움을 줄 수 있으며 변수로 판단된다. 엘로 평점 시스템은 홈/원정 여부, 과거 성적에 기반해서 만들어진 공식으로 인해 점수가 도출된다. 또한, 최근까지의 경기 결과를 바탕으로 누적되는 결과를 반영했다는 점에서 향후 경기 결과와의 연결성을 가질 가능성이 크다[15]. 따라서, 본 연구에서는 엘로 평점 시스템을 활용한 변수들을 추가하여 머신러닝 알고리즘에 반영하고 해당 변수들의 경기 결과에 대한 예측 기여도를 측정하여 변수의 유의성을 검증하고자 한다.

Ⅲ. 연구방법

3-1 연구의 표본

본 연구의 한국프로축구(K리그1) 경기 결과 예측을 위해 K리그1 정규리그 경기데이터를 활용하였다. K리그1 홈페이지에서는 2020시즌부터 “도움”, “오프사이드”, “프리킥”, “롱패스”, “숏패스”, “펀칭”, “골킥 성공” 등 기존보다 더욱 풍부한 자료를 게시함으로써 본 연구를 위한 한층 더 정밀한 분석이 가능해졌다. 따라서, 더욱 넓은 범위의 과학적이고 체계적인 데이터의 활용을 위해 본 연구에서는 2020~2023시즌의 경기기록을 활용하여 데이터 분석에 적용하였다.

코로나19의 영향으로 시즌이 중단된 2020시즌은 160경기의 양 팀 기록에 해당하는 320개의 데이터를 연구의 표본으로 설정하였고, 2023시즌의 경우 실증분석이 진행된 시점(~9월 3일)까지 348경기의 양 팀 기록에 해당하는 348개의 데이터를 연구의 표본으로 설정하였다. 본 연구대상 기간에 따른 표본 수는 표 1과 같이 정리될 수 있으며, 표 2는 분석이 진행된 기간 K리그1에 소속되었던 15개 팀의 홈/원정 경기 수와 참여 시즌을 나타낸 것이다.

K리그 공식 홈페이지(https://data.kleague.com)에서 제공하는 선수별 경기력 관련 상세 기록을 Python 3.10.9의 Selenium과 BeautifulSoup 라이브러리를 활용하여 수집하였다. 수집한 데이터 중 드리블, 패스, 경합 등 백분위 성공률

표 1. 연구대상 자료 측정 항목
Table 1. Measurement items

Season	Data Samples
2020	320 (160 Matches)
2021	456 (228 Matches)
2022	456 (228 Matches)
2023	348 (174 Matches)

표 2. 연구 표본 구단의 참가 경기 및 시즌 수
Table 2. Number of participating matches and seasons in the sample clubs

Team	Home	Away	Total	Season
Ulsan HD	66	65	132	4
Pohang Steelers	67	65	132	4
Daegu FC	65	67	132	4
FC Seoul	67	65	132	4
Jeonbuk Hyundai M	67	65	132	4
Incheon United	66	66	132	4
Gangwon FC	66	66	132	4
Suwon Samsung BW	67	66	132	4
Gwangju FC	46	48	94	3
Jeju United FC	51	54	105	3
Suwon FC	53	52	105	3
Seongnam FC	52	51	103	3
Gimcheon Sangmu	32	33	65	2
Daejeon Citizen	14	15	29	1
Busan IPark	13	14	27	1

형태로 제공되는 데이터에 대해서는 성공 개수와 성공률을 기반으로 “시도” 변수를 생성하였으며, 선수별 데이터를 기반으로 팀 단위의 데이터 프레임을 생성하기 위해 평균(“평점”), 합(“슈팅”, “유효 슈팅”, “오프사이드” 등), 고윳값(“심판”, “홈 여부”, “홈 승리”, “감독” 등)을 사용해 데이터를 전처리하였다.

나아가 본 연구에서는 국내 온라인 베팅 사이트인 배트맨(betman)에서 제공하는 프로토 승부식의 배당률 기록을 변수화하여 경기 결과 예측에 활용하기 위해 2020년부터 2023년까지의 “홈 팀 승리”, “무승부”, “어웨이 팀 승리”에 대한 K리그 배당률 데이터에 대해 체계적으로 수집하였다.

데이터의 구성상 2020년은 24회 차부터 75회 차까지 K리그1 경기가 존재했던 28회에 대한 배당률 데이터를 구축하였다. 2021년은 16회 차부터 96회 차까지 K리그1 경기가 존재했던 53회에 대한 배당률 자료를 수집하였다. 2022년은 15회 차부터 86회 차까지 K리그1 경기가 존재했던 42회에 대한 배당률 데이터를, 2023년은 24회 차부터 실증분석이 진행된 시점인 102회 차까지 K리그1 경기가 존재했던 33회에 대한 배당률 자료를 수집하였다.

3-2 연구변수

1) 경기 관련 / 경기 외적 연구변수

본 연구의 경기 관련 변수는 크게 경기 내에서의 통계 지표인 경기 관련 변수와 그 외의 경기 외적 변수로 구분할 수 있다. 경기 관련 변수는 선수 단위로 수집하였고, 경기 외적 변수는 팀 단위로 수집하였다. 표 3은 본 연구에서 사용한 120개의 경기 관련 변수를 나타내며, 표 4는 8개의 경기 외적 변수를 나타낸다.

표 3. 연구의 변수 - 경기 관련 변수
Table 3. Research variables - On-field variables

On-field Variables	
Full Time Team Goal	Rating
Shot	Shot On Target
Blocked Shot	Shot Off Target
Shot inside Penalty Area	Shot outside Penalty Area
Offside	Free Kick
Corner Kick	Throwing
Dribble Success	Dribble Attempt
Pass Success	Pass Attempt
Pass Success in Offense Area	Pass Attempt in Offense Area
Pass Success in Middle Area	Pass Attempt in Middle Area
Pass Success in Defense Area	Pass Attempt in Defense Area
Long-distance Pass Success	Long-distance Pass Attempt
Middle-distance Pass Success	Middle-distance Pass Attempt
Short-distance Pass Success	Short-distance Pass Attempt
Forward Pass Success	Forward Pass Attempt
Side Pass Success	Side Pass Attempt
Back Pass Success	Back Pass Attempt
Cross Success	Cross Attempt
Press-resistance	Key Pass
Duel Success on Ground	Duel Attempt on Ground
Duel Success on Air	Duel Attempt on Air
Tackle Success	Tackle Attempt
Clearance	Intercept
Pass Block	Ball Gain
Block	Ball Miss
Foul	Gain Foul
Yellow Card	Red Card
Goalie Punching	Goalie Catching
Kick Success by Goalie	Kick Attempt by Goalie
Saved Air-ball by Goalie	Attempt to Save Air-ball by Goalie
Full Time Team Score (Opp)	Rating (Opp)
Shot (Opp)	Shot On Target (Opp)
Blocked Shot (Opp)	Shot Off Target (Opp)
Shot inside Penalty Area (Opp)	Shot outside Penalty Area (Opp)
Offside (Opp)	Free Kick (Opp)

Corner Kick (Opp)	Throwing (Opp)
Dribble Success (Opp)	Dribble Attempt (Opp)
Pass Success (Opp)	Pass Attempt (Opp)
Pass Success in Offense Area (Opp)	Pass Attempt in Offense Area (Opp)
Pass Success in Middle Area (Opp)	Pass Attempt in Middle Area (Opp)
Pass Success in Defense Area (Opp)	Pass Attempt in Defense Area (Opp)
Long-distance Pass Success (Opp)	Long-distance Pass Attempt (Opp)
Middle-distance Pass Success (Opp)	Middle-distance Pass Attempt (Opp)
Short-distance Pass Success (Opp)	Short-distance Pass Attempt (Opp)
Forward Pass Success (Opp)	Forward Pass Attempt (Opp)
Side Pass Success (Opp)	Side Pass Attempt (Opp)
Back Pass Success (Opp)	Back Pass Attempt (Opp)
Cross Success (Opp)	Cross Attempt (Opp)
Press-resistance (Opp)	Key Pass (Opp)
Duel Success on Ground (Opp)	Duel Attempt on Ground (Opp)
Duel Success on Air (Opp)	Duel Attempt on Air (Opp)
Tackle Success (Opp)	Tackle Attempt (Opp)
Clearance (Opp)	Intercept (Opp)
Pass Block (Opp)	Ball Gain (Opp)
Block (Opp)	Ball Miss (Opp)
Foul (Opp)	Gain Foul (Opp)
Yellow Card (Opp)	Red Card (Opp)
Goalie Punching (Opp)	Goalie Catching (Opp)
Kick Success by Goalie (Opp)	Kick Attempt by Goalie (Opp)
Saved Air-ball by Goalie (Opp)	Attempt to Save Air-ball by Goalie (Opp)

* Opp: 상대편 기록을 나타내는 변수

표 4. 연구의 변수 - 경기 외적 변수
Table 4. Research variables - Off-field variables

Off-field Variables	
Match Date	Home Team
Away Team	Match Referee
Is Home	Home Win
Away Win	Team Coach

모든 경기 관련 변수와 경기 외적 변수 중 "홈 승리", "어웨이 승리"는 경기 이전의 선행적 예측을 위해 이전 경기들에 대한 평균치로 값을 대체하였다. 이때 k경기에 대한 평균을 산출하는 시계열 파라미터 k에 대해 정의하고, 자체 실험 결과 승부 예측력이 가장 뛰어났던 $k=7$을 본 연구에서는 활용하였다.

2) 엘로 평점 시스템 관련 변수

엘로 평점 시스템은 그 신뢰성과 범용성을 인정받아 많은 스포츠에서 선수나 팀에 대한 능력 평가로 그 기능이 확장되었다[15],[16]. 엘로 평점 시스템은 최근 경기에서 선수 혹은 팀의 성과를 해당 팀의 역량에 대한 동적인 평가(dynamic evaluation)를 가능하게 한다. 프로축구 경기에 있어서 시간의 흐름에 따라 경기와 그 결과가 발생하는 시계열적 특성은 엘로 평점 시스템을 프로축구팀의 경기력을 측정하는 지표로서 적합하다고 평가된다[18]. 프로축구 경기에서 경쟁하는 양 팀에 대한 엘로 평점 시스템은 해당 팀의 과거 성과에 해당하는 eq.(1), eq.(2)와 현시점 경기의 결과인 eq.(3), eq.(4)에 따라 결정된다.

$$E^H = \frac{1}{1+c^{\frac{R^H-R^A}{d}}} \tag{1}$$

$$E^A = 1 - E^H \tag{2}$$

$$S^H = \begin{cases} 1 & Home\ Win \\ 0.5 & Draw \\ 0 & Away\ Win \end{cases} \tag{3}$$

$$S^A = 1 - S^H \tag{4}$$

E^H와 E^A는 홈 팀과 원정팀에 대한 각 기대 성과 점수이며 이는 경기 시점에 대한 각 팀의 성과 점수 R^H와 R^A 및 평가 척도에 따른 가중치 상수 c와 d로 이뤄진다. 본 연구에서는 엘로 평점 시스템 관련 선행연구를 체계적으로 적극적으로 참고함과 동시에[16],[19],[20] 향후 연구의 확장성을 고려하여 가중치 상수 c와 d에 범용성 높은 수치인 10과 400을 각각 설정하였다. S^H와 S^A는 양 팀의 경기 결과에 평점을 의미한다. 추가로, 각 팀의 성과 점수 도출은 다음과 같다.

$$w = w_0(1+\sigma)^{\gamma} \tag{5}$$

$$R^{H\prime} = R^H + w(S^H - E^H) \tag{6}$$

w는 평점 수정 가중치로 eq.(5)와 같이 작용한다. 즉, 이는 경기 결과에 대한 엘로 평점의 수정 강도를 조작하는 변수로써 상세 구성은 다음과 같다. σ는 절대 골 득실에 대한 정보이며, w_0와 γ는 수정 가중 상수로 이 또한 엘로 평점 시스템 관련 선행연구는 물론 본 연구의 확장성을 고려하여 각각 10과 1이라는 범용성 높은 수치를 설정하였다[16].

이렇게 산출한 평점 수정 가중치 w를 eq.(6)과 같이 활용하여 실제 경기 결과와 엘로 평점 시스템상의 기대 결과 간의

차이를 통해 각 팀의 엘로 평점을 시즌 중 계속해서 갱신함으로써 팀에 대한 동적 평가를 구현한다.

본 연구에서 K리그1의 각 팀에 대한 엘로 평점 시스템은 2020시즌부터 측정이 시작되었으며 초깃값은 엘로 평점 시스템의 표준 점수인 1,500으로 설정했다. K리그1 소속 팀들에 대해 리그 일정을 참고하여 상대하는 팀과 경기 결과에 따른 엘로 평점 시스템을 계속해서 수정해 나갔으며[16],[19], 예측에 사용되는 양 팀의 엘로 평점은 해당 경기 직전 시점인 양 팀에 대한 엘로 평점을 활용하였다.

한편 승격과 강등이 존재하는 K리그 특성상 승격 팀 중 2020시즌 개막 이래로 처음 K리그1에 소속되는 팀에 대해서는 엘로 평점 시스템의 표준 초깃값인 1,500을, 2020시즌 이후 K리그1에 소속되었던 전적이 있는 승격 팀에 대해서는 K리그2 강등 확정 경기인 K리그1에서의 마지막 경기 결과가 반영된 엘로 평점을 부여하였다. K리그1 잔류 팀에 대해서는 이전 시즌 각 팀의 마지막 결과가 반영된 엘로 평점으로 새로운 시즌을 시작할 수 있도록 하였다. 이후 2023시즌 29라운드까지 같은 방식으로 팀별 엘로 평점을 산출하였다.

추가로, 엘로 평점 시스템 기반 회귀 모델을 사용하여 프로축구 경기 결과에 대한 사전 확률을 산출한 해외 선행연구를 참고하여[19], 엘로 평점 시스템 관련 파생변수를 만들어 경기 예측력을 증가시키고자 하였으며, 승/무/패 각각의 확률에 대한 산출 공식은 다음과 같다[16].

$$P_H = 0.448 + (0.0053 \times (E^H - E^A) \quad (7)$$

$$P_A = 0.245 + (0.0039 \times (E^A - E^H) \quad (8)$$

$$P_D = 1 - (P_H + P_A) \quad (9)$$

이때 P_H는 홈 팀의 승리 확률 (7), P_A는 어웨이 팀의 승리 확률 (8), P_D는 무승부 확률 (9)을 나타낸다. 이를 바탕으로 파생시킨 엘로 평점 시스템 기반의 추가 변수들은 표 5와 같다.

표 5. 엘로 평점 시스템 기반 파생변수
Table 5. Derived variables based on Elo rating system

Variable Name	Explanation
Elo	Elo rating of the home team
Elo_Away	Elo rating of the away team
Elo_WinRate	Winning rate of home team based on its elo rating
Elo_DrawRate	Draw rate of both teams based on their elo ratings
Elo_WinRate_Away	Winning rate of away team based on its elo rating
Elo_STD	Standard deviation of elo ratings of both teams

3) 예측 변수 선정 및 변수 중요도 확인

본 연구는 경기 관련 변수, 경기 외적 변수, 엘로 평점 시스템 기반 변수, 배당률 기반 변수를 포함하여 130개가 넘는 경기 관련 특징을 독립 변수화하였다. 그러나 이처럼 많은 변수는 모델의 분산을 과하게 증가시켜 과적합이 유발될 가능성이 크다[21]-[23]. 이는 변수 선정(feature selection)으로 해결할 수 있는데, 변수 선정은 두 가지 중요성을 내재한다고 볼 수 있다. 변수 선정으로 과적합을 방지함으로써 모델의 성능 향상을 유도할 수 있으며, 나아가 모델을 통해 생성된 데이터와 그 배경을 이해함으로써 더욱 깊은 모델에 대한 통찰을 확보할 수 있게 된다[23].

본 연구에서는 다변량 변수 중요도를 간편하고 빠르게 산출할 수 있음과 동시에 적은 표본 크기 및 고차원 데이터에도 강건하게 작동하며, 데이터의 변수 중요도를 폭넓게 수치화할 수 있는[24],[25] scikit-learn 라이브러리의 Rand- om Forest Classifier[26] 내 Feature Importance 기능을 활용하여 Gini Impurity Index를 기반으로[27] 각각의 변수가 경기 결과 예측에 미치는 중요도를 확인하였다.

이때 경기 결과 예측에 있어서 변수가 가지는 설명력이 소수점 둘째 자리까지 유효하지 않다면 해당 변수들을 예측에 활용하지 않는 것을 기준으로 변수 선정 프로세스를 정밀하게 갖추었다. 추가로, 엘로 평점 시스템 기반 변수들이 경기 결과 예측에 있어서 어느 정도의 변수 중요도를 가지는지 확인함으로써, 엘로 평점 시스템의 경기 결과 예측에 대한 기여도를 확인하였다.

3-3 $\mu+\lambda\sigma$ 의사결정론

본 연구에서는 더욱 높은 경기 예측 정확도를 확보하기 위해, 머신러닝 알고리즘을 통해 산출된 경기 예측 결과에 대하여 추가적인 표본 통계량 기반의 의사결정론을 개발하여 도입하였다. 개발한 의사결정론의 구체적인 원리는 다음과 같다.

(1) 예측 결과별 확률 산출: 경기 결과 예측에 대한 가장 좋은 성능평가 지표를 기록한 모델에 대해 표 6에서와 같이 경기별 "홈 팀의 승리 확률(p)", "무승부 확률(q)", "원정팀의 승리 확률(r)"을 predict_proba 모듈을 통해 산출한다.
(2) 경기별 통계량 산출: p, q, r의 평균 μ와 표준편차 σ를 산출한다.
(3) 예측 확정 및 포기: p, q, r의 최댓값이 $(\mu+\lambda\sigma)$ 이상일 경우, 해당 최댓값이 속한 범주로 예측을 확정하는 반면, p, q, r의 최댓값이 $(\mu+\lambda\sigma)$ 미만일 경우, 예측을 포기한다.
(4) 2023시즌 K리그1의 예측 대상 경기에 대해 반복: 남은 경기가 없어질 때까지 경기별로 (2)와 (3)의 단계를 반복한다.

표 6. 머신러닝 기반 경기결과 예측 양식
Table 6. Prediction form of the game result based on machine learning

Winning Rate of Home Team	Draw Rate of Both Team	Winning Rate of Away Team
p	q	r

예측가중치 변수 λ의 범위를 $\{\lambda|0 \leq \lambda \leq 2\}$로 정의하고 λ의 값의 변화에 따른 예측 정확도와 예측에 대한 의사결정 수행률을 비교 분석하여, 높은 예측 정확도를 갖는 동시에 높은 예측 의사결정 수행률을 갖는 최적의 λ를 찾아냄으로써 K리그 경기 결과 예측에 최적화된 예측가중치를 확정한 이후, $\mu+\lambda\sigma$ 시스템 기반의 2023 K리그1 경기 예측을 수행하였다.

3-4 머신러닝 알고리즘

1) Naive Bayes

본 연구에 적용된 머신러닝 알고리즘 중 하나인 나이브 베이즈 분류기는 기계학습과 데이터 분석에서 다양한 분류 작업에 사용되는 확률적 알고리즘이다. 이는 조건부 확률을 사용하여 사건의 확률을 갱신하는 베이즈 정리에 기반한다. 해당 알고리즘은 분류 측면에서 간결함, 효율성 및 효과적인 성능으로 인해 대중화되어 있으며, 대량의 특징 공간을 효율적으로 처리할 수 있다는 장점을 갖고 있다. 단순한 가정에도 불구하고, 해당 분류 알고리즘은 실질적 적용에 있어서 효과적인 성능을 보이기에 텍스트 분류, 의사결정 진단, 시스템 성과 관리 등 폭넓게 사용된다[28]-[30].

2) Logistic Regression

본 연구에 적용된 머신러닝 알고리즘 중 하나인 로지스틱 회귀는 기계학습 알고리즘 중 분류 작업에 사용되는 통계적인 방법으로, 데이터가 어떤 범주에 속할 확률을 예측한다. 일반적으로 종속변수가 두 개인 문제에 적용하는 이진 로지스틱 회귀 모델이 가장 많이 사용된다[31]. 이는 선형 방정식을 이용하여 입력 변수가 각 범주에 속할 확률 사이의 관계를 모형화하며, 시그모이드 함수를 통해 확률을 0에서 1 사이의 값으로 제한한다. 본 연구에서는 승/무/패의 다항 분류를 수행하기 위해 다항 로지스틱 회귀 모델(Multinomial Logistic Regression)을 활용하였다.

3) Light-GBM

본 연구에 적용된 머신러닝 알고리즘 중 하나인 Light-GBM 분류기는 경량화된 그래디언트 부스팅 기반 기계학습 알고리즘으로, 큰 규모의 데이터 세트나 고차원 특징 공간에서 효율적인 분류를 위해 설계된 빠르고 효과적인 알고리즘이다. 반면 작은 데이터 세트에서는 과적합 위험의 소지도 동반한다[32],[33]. 해당 알고리즘은 트리 기반의 알고리즘으로 구성된다. 기존의 알고리즘이 나무를 수평적으로 확장하는 반면, Light-GBM은 수직적으로 확장하며, 같은 데이터를 평가할 때 더 낮은 손실률을 보여준다[34].

4) Elastic Net

본 연구에 적용된 머신러닝 알고리즘 중 하나인 Elastic Net 분류기는 scikit-learn 라이브러리의 로지스틱 회귀 내에서 하이퍼 파라미터의 조정으로 구현될 수 있다. 로지스틱 회귀에 L1(LASSO) 및 L2(LIDGE) 정규화를 통합함으로써 Elastic Net은 특징 선택과 가중치 정규화의 이점을 결합한다. 이 결합은 하이퍼 파라미터 λ1과 λ2를 통해 조절되어 적절한 균형을 찾으며 이뤄지며, 그로 인해 다중공선성 위험이 있는 LASSO 모형과 다르게 두 종류의 벌점 부과(penalization)를 통해 다중공선성의 위험을 감소시킨다는 장점이 있다고 볼 수 있다[35]-[37].

5) Decision Tree

본 연구에 적용된 머신러닝 알고리즘 중 하나인 Decision Tree 분류기[38]는 트리 구조를 기반으로 주어진 입력에 따라 여러 조건을 고려해 나무 모양의 가지를 따라 이동하며 결과를 도출한다. 각 노드는 특정 조건에 대한 판단 기준이 되고, 각 조건에 관한 결과에 대응되며 모델은 데이터의 패턴을 학습하고 새로운 데이터에 대한 예측을 수행할 수 있게 된다. Decision Tree는 결과에 대한 해석이 비교적 쉬우며, 직관적인 구조를 갖추고 있어 결정 과정을 이해하기 쉽다는 특징을 갖는다[39]. 또한, 범주형과 수치형 데이터 모두에 적용할 수 있으며, 과적합을 방지하기 위해 최대 깊이 제한 및 최소 가지치기 등의 기법을 추가로 사용할 수 있다.

3-5 자료처리

본 연구에서는 K리그1의 경기 결과 예측을 위해 머신러닝 알고리즘의 적용은 물론 경기 관련/경기 외적 변수를 활용하여 팀별 시계열적 퍼포먼스의 변화를 경기 결과 예측에 사용하기 위해 다음과 같은 자료처리 과정을 거쳤다. 먼저, 시계열 평균 데이터는 시즌 종료와 다음 시즌 개막 사이의 선수와 감독 이적이 모델의 예측력을 저하하는 것을 방지하기 위해 시즌별 시계열 평균 기록이 다음 시즌에 영향을 미치지 않도록 모델 학습을 디자인하였다. 아울러, 시계열 평균의 대상이 되는 경기의 해당 팀과 상태 팀의 경기기록 평균에 대한 평균을 산정함에 따라, 팀 상성에 대한 기록을 추가하였다.

본 연구에 적용된 변수의 경우, 경기 외적 데이터 중 “날짜”는 데이터 프레임의 인덱스로 활용했으며 “홈 팀명”, “어웨이 팀명”, “심판”, “감독” 네 가지의 명목형 변수는 제거하여 최종적으로 271개의 독립변수에 대하여 변수 중요도를 확인하고 예측 변수 선정의 대상으로 고려하였다.

추가로 모든 모델별 알고리즘은 790경기의 각 팀을 1개의 샘플로 하는 1,580개의 경기데이터를 대상으로 학습 및 예측을 수행하였다. 이때 선수/감독의 이적 및 K리그 승격시스템에 따른 팀 변동 등을 고려하기 위해 무작위 분할이 아닌 시즌

별로 나누어 학습 데이터는 2020시즌부터 2022시즌까지, 예측 데이터는 2023시즌 8(k+1) 라운드부터 29라운드까지의 데이터로 선정하였다(학습/예측 데이터 비율- 0.79:0.21).

나아가, 모델별 알고리즘에 따른 예측 결과를 계산하여, 실제 결과와 예측 결과에 대한 예측 정확도(Accuracy), F1-Score를 계산하였다. 모델 간 비교 기준으로 적용한 정확도와 F1-Score의 도출은 다음과 같다.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$F1\,Score = 2 \times \frac{Recall \times \mathrm{P}recision}{Recall + \mathrm{P}recision} \tag{11}$$

F1-Score(11)의 경우, 정밀도(Precision)와 재현율(Recall)의 조화평균으로 0과 1 사이의 값의 범위를 가지며, 1에 가까울수록 좋은 분류기로 평가한다.

Ⅳ. 연구결과

4-1 변수 설명력 측정 결과

본 연구에서 수집 및 가공한 총 271개의 경기 결과 예측을 위한 독립변수 중 상위 12개 변수의 변수 중요도는 그림 1과 같다. 271개의 변수 중 가장 중요한 변수는 “배당률” 관련 변수로 나타났다. “홈 팀 승리에 대한 배당률”이 가장 중요한 변수로 선정되었으며, 변수 설명력은 약 9.65%를 보였다. 이어지는 중요 변수는 “어웨이 팀 승리에 대한 배당률”이었으며, 변수 설명력은 약 9.40%를 보였다.

엘로 평점 시스템 기반 파생변수의 설명력은 표 7에 나타난 바와 같다. 해당 파생변수 중 “엘로 평점 시스템 기반 무승부 확률” 변수를 제외하고는 경기 결과 예측에 대해 대체로 높은 변수 설명력을 보였다. “홈 팀의 엘로 평점”이 271개의 독립변수 중 세 번째로 중요한 것으로 확인되었으며 변수 설명력은 약 4.09%로 산출되었다. 전체 독립변수 중 다섯 번째와 여섯 번째로 중요한 변수들은 각각 “원정팀의 엘로 평점”과 “엘로 평점 기반 원정팀의 승리 확률” 변수였으며, 설명력은 각각 약 2.90%와 2.82%로 산출되었다.

표 7. 엘로 평점 시스템 기반 파생변수의 변수 설명력
Table 7. Feature importance of derived variables based on Elo rating system

Variable Name	Importance (%)
Elo	3/271 (4.09%)
Elo_Away	5/271 (2.90%)
Elo_WinRate_Away	6/271 (2.82%)
Elo_WinRate	12/271 (1.71%)
Elo_STD	79/271 (0.31%)
Elo_DrawRate	121/271 (0.00%)

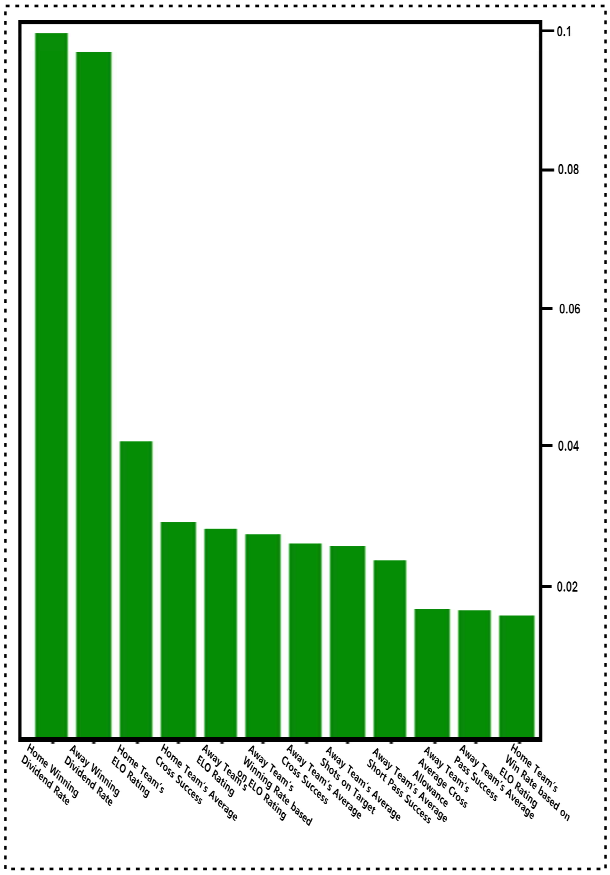


그림 1. 변수 중요도 도표
Fig. 1. Feature importance graph

전체 독립변수 중 “엘로 평점 기반 홈 팀의 승리 확률” 변수는 열두 번째 중요도를 가졌으며, 변수 설명력은 약 1.71%로 산출된 한편 “양 팀의 엘로 평점 간 표준편차”는 일흔아홉 번째 중요도를 가졌으며, 변수 설명력은 약 0.31%로 산출되었다. 반면, “엘로 평점 기반의 무승부 확률” 변수의 경우, 271개의 독립변수 중 151개의 변수와 함께 별수 설명력 0.00%로 소수점 둘째 자리 기준 유의하지 않은 변수로 확인되었다.

경기 결과 예측에 활용하는 변수(0.01% 이상의 변수 설명력을 가지는 변수) 중 최소 변수 설명력을 가지는 변수는 “홈 팀의 지상 경합 성공 횟수 평균” 변수였으며, 271개의 독립변수 중 120번째의 변수 중요도를 가졌으며, 약 0.12%의 변수 설명력을 가지는 것으로 확인되었다. 해당 변수는 변수 선정에 있어 분기점(Threshold)이 되어 해당 변수 이상의 변수 설명력을 가지는 총 120개의 독립변수가 경기 예측 변수로 선택되었다.

4-2 머신러닝 기반 경기 결과 예측 결과

본 연구는 K리그1 경기 예측에 가장 적합한 머신러닝 알고리즘을 탐색하고 엘로 평점 시스템의 경기 예측 변수로서의

유의성을 검증하기 위해 5개의 머신러닝 알고리즘과 변수 선정을 활용하여 총 5개의 2023시즌 K리그1 경기 결과 예측 모델을 생성하였다. 모델에서의 종속변수는 경기의 결과로 "승", "무", "패"로 분류된다.

표 8에서 나타난 머신러닝 알고리즘별 분류 예측 정확도(Accuracy)에 관한 분석을 살펴보면, Decision Tree 알고리즘을 적용했을 때 예측 정확도 0.48로 가장 성능이 우수했으며, Logistic Regression과 Elastic Net이 예측 정확도 0.46으로 그 뒤를 이었다. Light-GBM과 Naive Bayes의 경우 예측의 정확도가 0.41로 가장 낮은 성능을 기록했다. 분류 모델 성능평가 지표에서 Accuracy와 "승", "무", "패" 범주에 대해 비교적 균형 있게 높은 수치를 기록한 Decision Tree 모델이 경기 예측 의사결정론 도입을 위해 적합한 머신러닝 알고리즘으로 선택되었다.

표 8. 분류 모델 성능평가 지표 행렬
Table 8. Performance evaluation metrics

Algorithm	Accuracy	F1-Score		
		Win	Draw	Loss
Decision Tree	0.48	0.59	0.32	0.47
Logistic Regression	0.46	0.43	0.50	0.43
Elastic Net	0.46	0.41	0.52	0.41
Light-GBM	0.41	0.42	0.31	0.49
Naive Bayes	0.41	0.48	0.19	0.48

4-3 $\mu+\lambda\sigma$ 의사결정론 기반 경기결과 예측 결과

머신러닝 알고리즘별 분류 모델 성능평가에서 가장 우수한 성능을 보인 Decision Tree 모델을 기반으로 산출한 경기별 "홈 팀 승리 확률", "무승부 확률", "어웨이 팀 승리 확률"에 대해, $\mu+\lambda\sigma$ 시스템을 도입하고 λ의 변화에 따른 경기 결과 예측 정확도와 의사결정 수행률에 대한 변화는 그림 2와 같다.

λ가 증가할수록 의사결정 수행률(적색)은 점점 감소하다 일정 수치를 넘어가며 급격히 감소하고, 약 1.4를 초과하는 순간 더는 의사결정을 수행하지 못하고 0이 지속되는 것으로 나타났다. 경기 결과 예측 정확도(녹색)는 λ가 증가함에 따라 서서히 함께 증가하는 추세를 보이고, 의사결정 수행률이 0에 도달함에 따라 예측 정확도를 계산할 수 없게 됨이 확인되었다.

본 연구에서는 $\lambda=1.1$일 때, 의사결정 수행률 0.78과 더불어 예측 정확도 0.51을 기록하며 가장 균형 잡힌 K리그1 예측 가중치 변수로써 작용한다고 판단하여, 해당 수치로 2023시즌 K리그1 경기 결과 예측을 시도하였다.

본 연구의 실증분석 결과 2023시즌 K리그1의 8라운드부터 29라운드에 해당하는 132경기 중 경기 결과 예측에 대한 모델의 의사결정은 102경기에서 발생하였고(의사결정이 포기된 경기는 실제 홈 승리 7경기, 어웨이 승리 13경기, 무승부 10경기), 그중 52경기의 실제 결과에 대해 예측이 적중하여 약 51%의 예측 정확도로 K리그 경기 예측에 성공하였다.

본 연구의 머신러닝 알고리즘 및 의사결정론을 적용한 경기 예측 결과, 예측 모델의 의사결정이 일어난 102경기 중 홈 팀이 승리를 거둔 실제 44경기 중 예측 모델은 29경기를 승리, 8경기를 무승부, 7경기를 홈 팀의 패배로 예측하여 예측 정확도는 약 65.9%에 해당하였다. 무승부를 거둔 실제 32경기 중 예측 모델은 7경기에 대해 예측에 성공했고, 14경기를 홈 팀의 승리, 11경기를 홈 팀의 패배로 예측하여 예측 정확도는 약 21.9%로 산출되었다. 실제로 홈 팀이 패배한 26경기 중 예측 모델은 16경기를 패배로 예측했고, 8경기를 승리, 2경기를 무승부로 예측하여 약 61.5%의 예측 정확도를 보였다.

연구 결과, 본 연구에 적용된 머신러닝 알고리즘은 K리그1 경기 예측에 있어 "승리", "패배", "무승부"의 순서로 예측이 정확했으며, 무승부의 경우 특히 예측 정확도가 비교적 낮은

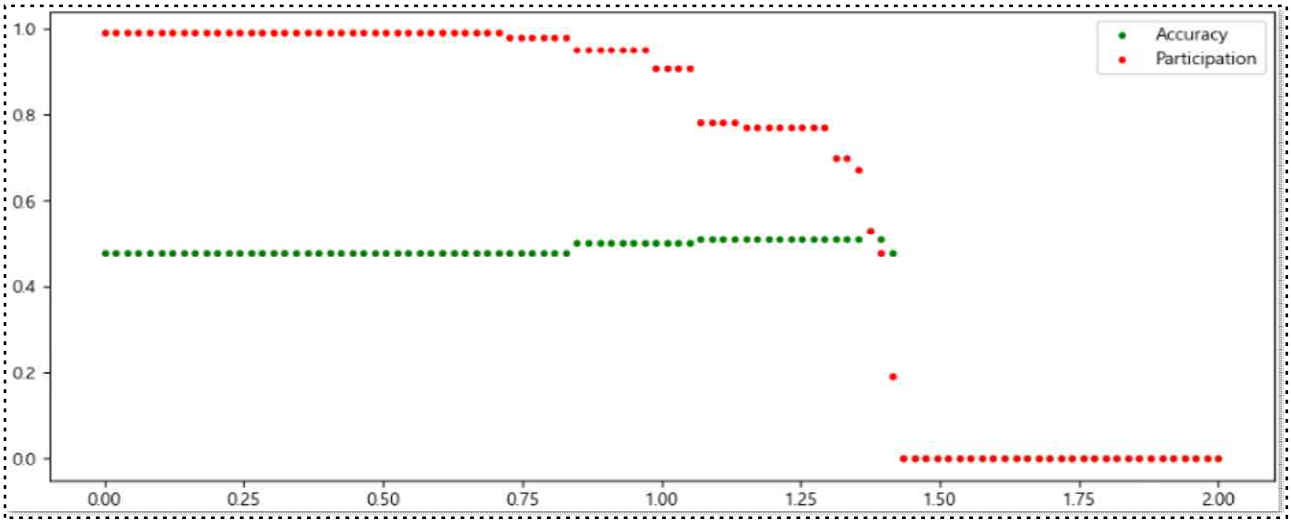


그림 2. λ의 변화에 따른 경기 결과 예측 정확도와 의사결정 참여율의 변화
Fig. 2. Changes in prediction accuracy of match result and decision-making participation rate by changes in λ

특징을 보였다. 표 9는 예측 모델의 예측이 일어난 K리그 2023시즌의 102경기를 예측한 결과와 실제 결과를 나타낸 것이다.

표 9. 의사결정이 일어난 K리그 2023시즌 예측 결과 및 실제 경기 결과 분포

Table 9. Distribution of prediction and actual match result of 2023 K-league (decision-made matches)

	Win (A)	Draw (A)	Loss (A)	Total
Win (P)	29	14	8	51
Draw (P)	8	7	2	17
Loss (P)	7	11	16	34
Total	44	32	26	102

* (P): 모델 예측 (Prediction)
* (A): 실제 결과 (Actual)

V. 논 의

본 연구는 머신러닝 알고리즘을 적용하여 한국프로축구 리그의 2023년 8라운드부터 29라운드에 해당하는 132경기 중 102경기를 5가지 알고리즘(Naive Bayes, Logistic Regression, Light GBM, Elastic Net, Decision Tree)을 기반으로 엘로 평점 시스템 관련 변수를 포함한 120개의 변수를 활용하여 경기 예측에 대한 실증분석을 진행하였다.

본 연구의 실증결과, “엘로 평점 기반 무승부 확률” 변수를 제외한 5가지의 엘로 평점 시스템 관련 변수들이 경기 결과에 대해 유의한 예측 변수로 선택될 뿐 아니라 높은 변수 중요도를 보인다는 것이 나타났다. 이에 따라 엘로 평점 시스템의 축구 경기 예측에 있어서 가지는 유의성을 확인하였다. 더 나아가 5가지 머신러닝 알고리즘 중 가장 높은 예측 정확도를 보인 Decision Tree 기반 예측 모델에 $\mu+\lambda\sigma$ 의사결정 시스템을 도입하여 최종 예측 정확도 0.51을 산출하였다.

이상의 분석과 결과를 바탕으로 본 연구는 다음과 같은 이론적 시사점에 대해 논의할 수 있다. 첫째, 본 연구의 머신러닝 알고리즘과 엘로 등급 관련 변수의 적용을 통한 실증분석으로 아직 경기 결과 예측이 충분히 이루어지지 않은 국내 프로축구 종목에서 국내 최초로 결과를 제시하였다. 전술하였듯이, 프로축구는 상세한 경기기록이 축적되기 어려운 플레이 특성, 적은 득점의 발생과 단순한 승패 예측을 넘어 무승부를 예측해야 한다는 어려움으로 인해 다른 프로스포츠 종목과 비교할 때 예측 정확도를 높이는 데 한계가 존재하였다. 그로 인해 인공지능의 적용과 스포츠 애널리틱스의 발전에도 불구하고 현시점까지 프로축구 경기 결과에 대한 예측이 충분히 이루어지지 않은 실정이다[9],[40].

이러한 연구의 공백은 국내 K리그1의 경우 더욱 심각하다고 볼 수 있다. 한국프로축구는 프로야구와 더불어 가장 큰 산업적 가치를 지닌 종목임에도 불구하고 경기데이터의 축적이 미비한 상태이며 선행적 예측이 아닌 독립변수와 종속변수가 동시에 활용되는 “예측” 연구조차 발견하기 어려운 실정이다. 본 연구는 이러한 연구의 공백을 해결하고, 한국프로축구의 스포츠 애널리틱스 적용과 발전을 위해 경기가 시작되기 전 기존에 존재하는 경기데이터의 조합만으로 경기 결과를 실시간적으로 예측하고 유의미한 결과를 도출하였다는 점에서 이론적 의의를 지닌다.

둘째, 본 연구에서는 국내 프로스포츠 최초로 엘로 평점 시스템을 적용해 경기 결과를 예측하였다. 엘로 평점 시스템이 국제축구연맹(FIFA: Fédération Internationale de Football Association), 미국프로야구(MLB: Major League Baseball), 미국프로농구(NBA: National Basketball Association), 내셔널 풋볼 리그(NFL: National Football League), 내셔널 하키 리그(NHL: National Hockey League) 등 다양한 종목의 협회나 기구에서 선수나 팀의 등급을 측정하기 위해 사용하지만, 머신러닝 알고리즘을 적용하여 경기 결과를 더욱 정밀하게 예측하기 위해 이를 사용한 경우는 찾기 어렵다.

특히, 국내 프로스포츠 종목의 경기 결과를 예측하기 위해서 엘로 평점 시스템과 관련한 변수들을 생성하여 이를 반영하면 실제 예측 성능이 개선된다는 것을 확인한 연구는 존재하지 않는 것으로 보인다. 본 연구에서는 경기 결과에 대한 선행적 예측 연구가 존재하지 않는 한국프로축구 경기기록에 엘로 평점 시스템 관련 변수들을 추가하여 경기 결과 예측에 적용하게 되면 예측 성능이 개선된다는 것을 다양한 머신러닝 알고리즘을 통해 실증적으로 증명함으로써 후속 연구에서 활용할 수 있는 좋은 길잡이가 될 것으로 기대할 수 있다.

셋째, 본 연구에서는 프로축구 종목의 경기 결과 예측에 어려움을 겪게 되는 무승부에 대하여 표본 통계량 기반의 의사결정론을 개발하여 도입함으로써 해결책을 제시하였다. 경기 데이터를 바탕으로 프로축구 경기의 결과를 예측한 기존의 선행연구에서는, 무승부 경기를 예측한 비율이 승패를 예측해서 적중하는 비율보다 낮았다[9]. 즉, 프로축구 경기의 결과 예측에 어려움을 제공하는 가장 큰 요인 중 하나는 승/패가 아닌, 승/무/패로 이루어진 형태의 결과라고 볼 수 있으며, 실제 무승부를 정확하게 예측할 가능성이 승패에 비해 현저하게 낮아지는 현상이 발견되었다. 이를 해결하기 위해 머신러닝 기반 다중 분류에 대한 범주별 발생 확률과 그 표본 통계량 및 가중치를 활용하여 예측에 반영했다는 점에서 기존의 연구에서 무승부 예측 정확도가 낮아진다는 한계를 개선할 수 있는 실마리를 제공하였다.

아울러, 본 연구는 다음과 같은 점에서 실무적 시사점을 지닌다. 먼저 스포츠 관련 미디어나 이벤트 관계자들은 머신러닝 알고리즘에 기반한 경기 결과 예측을 활용한 다양한 산업적 가치를 창출하고 마케팅 효과를 극대화함으로써 팬의 호기심을 자극할 수 있다. 현재까지 한국프로축구 리그 경기 결과를 사전에 예측하고 이에 대한 정보를 제공하거나, 선수의 활약상을 데이터 분석에 기반하여 제공하는 사례는 실무적으로 찾아보기 힘들다. 따라서, 프로축구 팬들의 전반적인 분석 능

력 수준과 데이터 기반의 자료에 대한 관심도가 매우 높아진 점을 고려하여 다른 프로스포츠 종목과 같이 데이터에 기반한 경기 결과의 선행적 예측이나 깊이 있는 분석자료를 제공하는 노력이 필요할 것으로 보이며, 본 연구는 이에 대한 기본적인 아이디어를 제공했다.

무엇보다 본 연구의 활용은 스포츠 베팅과 같은 새로운 스포츠 산업의 활성화를 위해 실무적인 활용도가 높을 것으로 보인다. 세계적으로 스포츠 베팅 산업은 그 어느 분야와 견주어도 가장 빠르게 성장하고 있는 영역 중 하나이며, 새로운 산업이라는 특성상 본 연구에서 파생될 수 있는 인공지능 기반의 스포츠 경기 예측으로 인한 확장 효과 역시 뛰어날 것으로 예상된다. 또한, 스포츠 베팅에 참여한 팬은 그렇지 않은 경우보다 더 높은 수준의 몰입을 보인다는 점에서 본 연구의 결과를 적용한 실무적 활용성에 대해 고민해 볼 필요가 있다[41]-[43].

프로스포츠 경기데이터를 바탕으로 경기 결과를 선행적으로 예측할 수 있는 내재적 역량을 지닌 기업이나 정부, 협회, 구단, 미디어의 노력이 그 중요도에 비해 눈에서 뛰지 않는 실정이다. 이러한 측면에서 볼 때, 본 연구에서는 한국프로축구리그를 포함한 프로스포츠의 발전과 산업적 확장을 위해 경기 결과의 선행적 예측이 실무적인 차원에서 적용될 필요성이 매우 높다고 판단하며, 본 연구의 실증결과는 이를 위한 하나의 이정표를 제공함으로써 스포츠 산업 확장에 기여도가 있을 수 있다고 판단된다.

Ⅵ. 결론 및 제언

본 연구는 엘로 등급과 머신러닝 알고리즘을 적용하여 한국프로축구 경기데이터를 바탕으로 경기 결과를 선행적으로 예측한 최초의 연구라는 점에서 학술 및 실무적 의의를 지니며, 향후 연구의 이정표를 제공하였다는 점에서 그 가치를 지닌다. 향후 연구에서는 다음과 같은 점을 고려하여 연구를 더욱 첨예하게 확장함으로써 인공지능 알고리즘을 적용한 스포츠 애널리틱스 분야의 연구를 발전시킬 수 있을 것으로 기대한다.

첫째, 분류 기반 이외의 알고리즘을 사용하여 경기 결과 예측을 시도할 필요가 분명 존재한다. 승/무/패로 이루어진 경기 결과를 예측하기 위해서 본 연구에서는 분류 기반의 알고리즘을 적용했다. 하지만, 일부 선행연구에서는 승/무/패로 이루어진 경기 결과가 최종적으로 예측하고자 하는 변수라 하더라도 이를 도출하는 과정에서 득점, 실점, 득실 마진 등 점수와 관련한 요인을 활용하여 경기 결과 예측이 가능하다는 측면에서 보고하였다. 이런 방법은 축구에서도 승패 예측에 사용할 수 있을 것으로 예상되며, 향후 연구에서는 이를 활용하여 분석함으로써 예측 성능 향상에 있어 유의미한 결과를 도출할 수 있을 것으로 보인다.

둘째, 무승부 예측의 정밀도를 높이기 위한 체계적인 추가 연구의 필요성이 존재한다. 프로축구 경기는 농구나 야구 종목과 같은 다른 프로스포츠 종목과 비교하여 점수가 적게 나는 종목의 특성상 무승부의 경우의 수가 존재한다. 이러한 점에서 무작위로 승/패를 예측한다면, 1/2에 근접하는 확률을 내재한 다른 프로스포츠 종목의 경기 예측과는 달리 프로축구는 확률적으로 1/3의 적중 가능성을 지닌다는 점에서 경기결과 예측이 무승부의 존재 여부로 상대적으로 낮아지게 된다. 본 연구에서도 역시 엘로 평점 시스템 변수 및 예측에 대한 의사결정론을 개발하여 일정 부분 기존의 선행연구를 확장하고 고도한 측면이 있지만, 여전히 승/패와 비교해 무승부의 예측 정확도는 낮은 것으로 나타났다. 이에 무승부를 예측하기 위한 추가적인 통계 및 기계학습을 통한 분석 및 의사결정 고도화에 관한 연구가 이루어진다면, 보다 이론·실무적으로 큰 가치를 지닌 연구가 이루어질 것으로 기대된다.

셋째, 머신러닝 알고리즘을 적용한 선행적 프로스포츠 경기 결과 예측에 있어 본 연구에서 살펴본 바와 같이 엘로 등급 등을 포함한 더욱 의미 있는 변수를 중심으로 알고리즘 모형을 간결하게 구성하는 노력이 필요하다. 일반적으로 종속변수를 도출하기 위한 변수의 투입이 많아질수록 모델의 전반적인 설명량은 늘어날 수 있지만, 해당 모델이 예측의 관점에서 더욱 중요한 의미를 지니려면 간명성 역시 중요한 요인이 된다.

특히, 연구자료를 활용하여 경기 결과에 중요한 요인을 중심으로 훈련을 진행하거나 이에 맞는 선수를 영입하고자 하는 현장 관계자들은 경기에 미치는 영향력이 큰 요인을 발견함으로써 이에 맞는 선수구성 혹은 전략을 선택하는 것이 중요하다. 향후 연구에서는 변수를 비교적 간결하게 정리하여 알고리즘 모형화에 대한 간명성을 확보한다면, 이론적으로는 물론 실무적으로도 더욱 유의미한 연구 결과를 제시할 수 있을 것으로 기대한다.

참고문헌

[1] K. Schwab, *The Fourth Industrial Revolution*, New York, NY: Crown Business, 2016.

[2] B. C. Alamar, *Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers*, New York, NY: Columbia University Press, 2013.

[3] B. S. Baumer, G. J. Matthews, and Q. Nguyen, “Big Ideas in Sports Analytics and Statistical Tools for Their Investigation,” *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 15, No. 6, e1612, November/December 2023. https://doi.org/10.1002/wics.1612

[4] G. Fried and C. Mumcu, *Sport Analytics: A Data-Driven Approach to Sport Business and Management*, London, UK: Taylor & Francis, 2016.

[5] M. Troilo and A. Bouchet, “Professional Sports Organizations and Business Analytics: Monopoly Power vs Debt Financing,” *Journal of Applied Sport Management*, Vol. 14,

No. 4, pp. 15-21, 2022. https://doi.org/10.7290/jasm14o8i1

[6] P. Kim and S. H. Lee, "The Application of Big Data Analysis in Comparison of Machine Learning Algorithms to Predict Korean Professional Basketball League Team Results," *The Korean Journal of Physical Education*, Vol. 62, No. 2, pp. 263-277, March 2023. http://doi.org/10.23949/kjpe.2023.3.62.2.19

[7] P. Kim and S. H. Lee, "The Final Ranking Prediction of the Korean Professional Basketball League Using Machine Learning Algorithms: A Sports Analytics Perspective," *The Korean Journal of Measurement and Evaluation in Physical Education and Sports Science*, Vol. 25, No. 2, pp. 103-115, June 2023. http://doi.org/10.21797/ksme.2023.25.2.008

[8] Y.-J. Seo, H.-W. Moon, and Y.-T. Woo, "A Win/Lose Prediction Model of Korean Professional Baseball Using Machine Learning Technique," *Journal of the Korea Society of Computer and Information*, Vol. 24, No. 2, 17-24, February 2019. https://doi.org/10.9708/jksci.2019.24.02.017

[9] P. Kim, S. H. Lee, and S. S. Jeon, "A Study on the Prediction and Evaluation of Keirin Competition Rankings Using Machine Learning Application," *Korean Journal of Sport Management*, Vol. 28, No. 2, pp. 76-94, April 2023. https://doi.org/10.31308/KSSM.28.2.76

[10] J. H. Yi and S. W. Lee "Prediction of English Premier League Game Using an Ensemble Technique," *KIPS Transactions on Software and Data Engineering*, Vol. 9, No. 5, pp. 161-168, May 2020. https://doi.org/10.3745/KTSDE.2020.9.5.161

[11] P. Kim, "Predicting the Outcome of Korean Professional Basketball Games and Applying Sports Betting Using Artificial Intelligence Algorithms," *The Korean Journal of Physical Education*, Vol. 62, No. 5, pp. 339-361, September 2023. https://doi.org/10.23949/kjpe.2023.9.62.5.23

[12] F. Thabtah, L. Zhang, and N. Abdelhamid, "NBA Game Result Prediction Using Feature Analysis and Machine Learning," *Annals of Data Science*, Vol. 6, No. 1, pp. 103-116, March 2019. https://doi.org/10.1007/s40745-018-00189-x

[13] A. Zimmermann, S. Moorthy, and Z. Shi, "Predicting College Basketball Match Outcomes Using Machine Learning Techniques: Some Results and Lessons Learned," arXiv:1310.3607, October 2013. https://doi.org/10.48550/arXiv.1310.3607

[14] P. Kim, S. S. Jeon, and S. H. Lee, "The Application of Machine Learning Algorithms to Predict English Premier League Match Results," *The Korean Journal of Physical Education*, Vol. 62, No. 4, pp. 337-353, July 2023. https://doi.org/10.23949/kjpe.2023.7.62.4.24

[15] A. E. Elo, *The Rating of Chessplayers: Past & Present*, 2nd ed. New York, NY: Ishi Press International, 2008.

[16] Y. Ren and T. Susnjak, "Predicting Football Match Outcomes with eXplainable Machine Learning and the Kelly Index," arXiv:2211.15734, November 2022. https://doi.org/10.48550/arXiv.2211.15734

[17] P. Robberechts and J. Davis, "Forecasting the FIFA World Cup - Combining Result- and Goal-Based Team Ability Parameters," in *Proceedings of the 5th International Workshop on Machine Learning and Data Mining for Sports Analytics (MLSA 2018, Co-located with ECML/PKDD 2018)*, Dublin, Ireland, pp. 16-30, September 2018. https://doi.org/10.1007/978-3-030-17274-9_2

[18] P. K. Jain, W. Quamer, and R. Pamula, "Sports Result Prediction Using Data Mining Techniques in Comparison with Base Line Model," *Opsearch*, Vol. 58, No. 1, pp. 54-70, March 2021. https://doi.org/10.1007/s12597-020-00470-9

[19] L. M. Hvattum and H. Arntzen, "Using ELO Ratings for Match Result Prediction in Association Football," *International Journal of Forecasting*, Vol. 26, No. 3, pp. 460-470, July-September 2010. https://doi.org/10.1016/j.ijforecast.2009.10.002

[20] G. Angelini, V. Candila, and L. De Angelis, "Weighted Elo Rating for Tennis Match Predictions," *European Journal of Operational Research*, Vol. 297, No.1, pp. 120-132, February 2022. https://doi.org/10.1016/j.ejor.2021.04.011

[21] M. A. Babyak, "What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models," *Psychosomatic Medicine*, Vol. 66, No. 3, pp. 411-421, May 2004.

[22] D. M. Hawkins, "The Problem of Overfitting," *Journal of Chemical Information and Computer Sciences*, Vol. 44, No. 1, pp. 1-12, January 2004. https://doi.org/10.1021/ci0342472

[23] Z. Zhang, "Too Much Covariates in a Multivariable Model May Cause the Problem of Overfitting," *Journal of Thoracic Disease*, Vol. 6. No. 9, pp. E196-E197, September 2014. https://doi.org/10.3978/j.issn.2072-1439.2014.08.33

[24] E. Scornet, G. Biau, and J.-P. Vert, "Consistency of Random Forests," *The Annals of Statistics*, Vol. 43, No. 4, pp. 1716-1741, August 2015. https://doi.org/10.1214/15-AOS1321

[25] G. Biau and E. Scornet, "A Random Forest Guided Tour," *Test: An Official Journal of the Spanish Society of Statistics and Operations Research*, Vol. 25, No. 2, pp.

197-227, June 2016. https://doi.org/10.1007/s11749-016-0481-7

[26] L. Breiman, "Random Forests," *Machine Learning*, Vol. 45, No. 1, pp. 5-32, October 2001. https://doi.org/10.1023/A:1010933404324

[27] H. Tyralis and G. Papacharalampous, "Variable Selection in Time Series Forecasting Using Random Forests," *Algorithms*, Vol. 10, No. 4, 114, October 2017. http://doi.org/10.3390/a10040114

[28] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," *Machine Learning*, Vol. 29, No. 2-3, pp. 103-130, November 1997. https://doi.org/10.1023/A:1007413511361

[29] T. M. Mitchell, Artificial Neural Networks, Carnegie Mellon University, Pittsburgh: PA, Machine Learning 10-701, February 2010.

[30] J. L. Hellerstein, T. S. Jayram, and I. Rish, "Recognizing End-User Transactions in Performance Management," in *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00)*, Austin: TX, pp. 596-602, July-August 2000.

[31] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2013.

[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, ... and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, Long Beach: CA, pp. 3149-3157, December 2017.

[33] E. Al Daoud, "Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset," *International Journal of Computer and Information Engineering*, Vol. 13, No. 1, pp. 6-10, 2019. https://doi.org/10.5281/zenodo.3607805

[34] D. Yazbek, J. S. Sibindi, and T. L. Van Zyl, "Deep Similarity Learning for Sports Team Ranking," arXiv:2103.13736, March 2021. https://doi.org/10.48550/arXiv.2103.13736

[35] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY: Springer, 2009.

[36] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, pp. 267-288, 1996. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[37] H. Zou and T. Hastie, "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Vol. 67, No. 2, pp. 301-320, April 2005. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[38] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, Vol. 1, No. 1, pp. 81-106, March 1986. https://doi.org/10.1007/BF00116251

[39] Y. Song and Y. Lu, "Decision Tree Methods: Applications for Classification and Prediction," *Shanghai Archives of Psychiatry*, Vol. 27, No. 2, pp. 130-135, April 2015. https://dx.doi.org/10.11919/j.issn.1002-0829.215044

[40] R. Baboota and H. Kaur, "Predictive Analysis and Modelling Football Results Using Machine Learning Approach for English Premier League," *International Journal of Forecasting*, Vol. 35, No. 2, pp. 741-755, April-June 2019. https://doi.org/10.1016/j.ijforecast.2018.01.003

[41] Grand View Research, Sports Betting Market Size, Share & Trends Analysis Report by Platform, by Betting Type (Fixed Odds Wagering, Exchange Betting, Live/In-Play Betting, eSports Betting), by Sports Type, by Region, and segment Forecasts, 2023-2030, Author, San Francisco: CA, GVR-4-68039-539-7, February 2023.

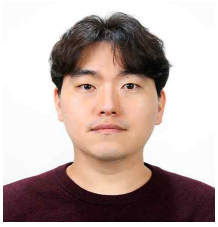
[42] H. Lopez-Gonzalez, F. Guerrero-Solé, A. Estévez, and M. Griffiths, "Betting is Loving and Bettors are Predators: A Conceptual Metaphor Approach to Online sports Betting Advertising," *Journal of Gambling Studies*, Vol. 34, No. 3, pp. 709-726, September 2018. https://doi.org/10.1007/s10899-017-9727-x

[43] Vantage Market Research, Sports Betting Market Size & Share to Surpass USD 129.3 Billion by 2028, Author, Washington, DC, November 2022.

김필수(Philsoo Kim)

2013년 : 연세대학교 대학원(경영학석사)
2022년 : 한양대학교 대학원(스포츠산업학박사)

2014년～2016년: 한국연구재단 Global PhD Fellow
2015년～2017년: 경기대학교 외래교수
2022년～현　재: 한국스포츠경영전략연구원 원장
※관심분야 : 스포츠경영, 스포츠 애널리틱스, 인공지능, 프로스포츠 등



이상현(Sang Hyun Lee)

2011년 : 아주대학교 대학원(경영학석사)
2024년 : 아주대학교 대학원(경영학박사)

2017년～2018년: 아주경영연구소 연구원
2019년～2022년: G. Lab 연구원
2023년～현　재: 한국스포츠경영전략연구원 부원장
※관심분야 : 스포츠 애널리틱스, 감독 역량, 리더십, 조직문화

서재현(Jae Hyun Seo)

2023년 : 한양대학교(체육학사, 공학사)
2024년 : 한양대학교 대학원(공학석사 과정)

2023년～2023년: 에이치앤컨설팅 연구원
2024년～현　재: 품질인텔리전스 연구실 연구원
※관심분야 : 스포츠애널리틱스, 생성형인공지능, 설명가능인공지능,
기술 혁신, 디지털 전환