KSME

http://doi.org/10.21797/ksme.2022.24.4.081

ISSN 1229-4225
eISSN 2671-9134

한국체육측정평가학회지. 제24권 4호, 2022, 81-91

축구의 경기 결과 예측을 위한 머신러닝 기법 비교*

최형준** 단국대학교

국문초록

본 연구는 축구 경기 결과의 예측을 위하여 사용되는 머신러닝 기법의 예측 성능을 비교하는 것을 주된 목적으로 두었다. 본 연구의 대상은 2013-2014시즌부터 2019-2020시즌까지 영국 EPL에서 개최된 경기(n=2,660)이었으며, 예측에 사용된 변인은 홈팀을 기준으로 하여 총 26개 독립변인과 1개 종속변인으로 선정하였다. 본 연구를 위해서 사용한 머신러닝 기법은 총 7가지로서, 로지스틱 회귀분석, 선형판별분석, 인공신경망 모델, 딥러닝 모델, 서포트벡터머신, 나이브 베이즈 모델, XGBoost 모델이 비교대상이 되었다. 모델의 평가방법은 Cohen's Kappa, 정확도(accuracy), 민감도(sensitivity), 특이도(specificity), 정밀도(precision), 재현율(recall), F1 점수로 비교하였다. 본 연구에서 축구 경기 결과를 예측하는 데 가장 뛰어난 예측 성능을 나타낸 것은 선형 판별 분석이었으며, 서포트벡터머신과 XGBoost 모델도 높은 F1 score를 나타내었다.

주요어: 축구 경기 결과 예측, 스포츠분석, 예측 성능 지표 비교

Ⅰ. 서론

스포츠경기분석은 체육측정평가 분야에서 전통적으로 연구되어 온 자료의 신뢰도와 타당도에 무게를 두고 다양한 연구 방법을 적용하여 해석하고 있다(최형준, 엄한주, 2020). 스포츠 경기에서 나타나는 경기력을 기술하기 위해 스포츠 경기에서 일어나는 주된 사건이나 내용을 체계적으로 관찰하고 기록하는 데 중점을 두고 있으며, 경기 중 나타나는 경기력이 실제 경기력이라는 것을 전제로 연구가 지속되고 있다. 실제로 스포츠 경기 중 관찰될 수 있는 다양한 사건과 현상을 체계적인 관찰 방법(systematic observation)을 통해 기록하고 분석하는 것은 관찰에 대한 객관성을 확보하는 데 중점을 두고 있으며, 다양한 학문 분야와의 융합이 가능하다(최형준, 김주학, 2006). 스포츠경기분석 분야에서는 다양한 학문 분야와의 융합을 통해서 인공지능 기법(artificial intelligence techniques)이나 데이터사이언스(data science) 분야에서 다루는 기법을 적용하고 있으며, 이에 관련한 연구도 늘어나고 있는 추세이다(최형준, 2020). 실제로 데이터사이언스의 범위는 기존의 통계적 접근을 비롯해서, 공학적 접근, 인공지능의 적용, 수학적 접근, 그리고 해당 분야에 대한 도메인 지식(knowledge of domain)을 융합하는 것을 필요로 한다(Shi et al., 2014). 여기에서 도메인 지식이란

* 이 연구는 2021학년도 단국대학교 대학연구비 지원으로 연구되었음(R-2021-00791).
** 교신저자 최형준(chj2812@dankook.ac.kr)
경기도 용인시 수지구 죽전로 152, 단국대학교

해당 분야에 관련된 지식을 말하며, 스포츠 분야의 도메인 지식인 경우에는 스포츠 분야에서 다루는 지식에 근거하여 데이터를 수집, 분석, 해석 및 표현하는 것이 중요하다는 것을 의미한다. 이러한 측면에서 스포츠경기분석에서 다루는 자료의 분석과 해석은 도메인 지식 내에서 설명되어야 되는 것이 타당하다.

스포츠경기분석 분야에서는 경기의 내용을 기록한 공식기록을 활용하여 경기 결과를 예측하고자 노력해왔다(이재현, 이수원, 2020; 이해용, 2012; 최형준, 이윤수, 2019; 홍종선, 정민섭, 이재형, 2010; Han et al., 2022). 축구 경기에 대한 예측은 경기 결과(Outcome)를 예측하는 방향으로 연구되어왔다. 경기 결과는 성공적인 내용과 그렇지 않은 내용과의 차이를 통해 승자와 패자를 구분하는 요소가 되며, 승자와 패자의 구분을 결정하는 요인으로 인식되어왔다(김주학, 최형준, 2015). 승자와 패자의 구분은 경기 결과에 의존하여 결정되는 결과이나, 그렇다고 해서 패자의 경기력이 상대적으로 낮거나 부족하다고 해석하는 것은 무리가 있다. 스포츠 경기의 결과는 상대하는 선수 혹은 팀이 어떠한 경기력 수준에 있는지에 따라서 상이하다(이범수, 최형준, 2015).

스포츠 경기의 결과는 상대성을 고려하여 예측해야 한다(김종원, 최형준, 2021). 스포츠 경기 결과의 예측에 대한 적중률을 높이기 위해서 고려해야 하는 다양한 제반 사항 중에서 경기내용에 대한 차이를 고려하는 것은 상대적인 경기력 수준에 대한 가변적 조건을 고려해야만 한다. 기존 예측에 활용되어왔던 인공지능 기법들은 학습자료에 결과를 포함하여 학습시키고, 이를 기반으로 하는 수리적 범위와 분산 등을 토대로 결과 데이터와의 차이를 살펴봄으로써 예측의 가능성을 살펴보았다. 이러한 방법에 있어서, 승리와 패배와 같이 이분적 경기 결과에 대한 예측은 팀이나 선수의 경기력 향상을 위한 현장 중심적 자료로 활용하는 데 한계가 있다(김주학, 최형준, 2015).

축구 경기는 득점 하나로 경기 결과가 결정되는 특징을 지닌다. 타 스포츠 종목 중에서는 기록에 의해 순위가 좌우되거나, 점수의 차이가 크기 때문에 승패가 결정되는 종목이 있는 반면에 축구 경기는 득점 하나로 인해 승패가 좌우된다(최형준, 이윤수, 2019). 축구 경기에서 득점의 차이는 분석하고자 하는 팀을 기준으로 양수(+)와 음수(-)가 될 수 있다. 분석하고자 하는 팀의 득점 차이가 0인 경우에는 상대하는 팀과 비교하여 득점에는 차이가 없으므로 비기는 결과를 나타낸다. 즉, 득점의 차이는 경기 결과를 승리와 패배와 같은 이분적 자료를 도출하는 중요한 단서를 제공한다. 승리한 팀이 패배한 팀에 비해서 득점을 많이 한 것은 사실이다. 축구 경기는 상대하는 팀의 경기력에 따라서 득점의 차이가 크게 나타날 수도 있고, 적게 나타날 수도 있다. 축구의 경기력에 대한 상대성(relativity)은 상대하는 팀과의 득점 차이에 따라서 해석될 수 있으며, 상대적 경기 결과의 대표적인 자료인 득점의 차이를 분석하고 예측하는 것은 상대성 지표(relativity index)의 기준을 설정하는 데 도움을 준다. 따라서 본 연구는 축구 경기 결과를 예측하는 데 있어서 홈팀을 기준으로 승리, 패배, 비김의 정도를 파악하고, 경기 결과의 예측에 사용되는 머신러닝 기법을 비교하는 데 주된 목적을 두었다.

Ⅱ. 연구방법

1. 연구대상 자료

축구 경기 결과의 예측을 위한 머신러닝 기법을 탐색하기 위하여 2013~14시즌부터 2019~20시즌(7년간)까지 영국 프리미어리그(English Premier League)에서 진행된 2,660경기를 연구의 대상으로 선정하였다. 〈표 1〉은 본 연구 대상이 된 시즌별 경기 수와 참가팀 수를 나타낸 것이다.

〈표 2〉는 본 연구의 대상이 된 경기에 참여한 팀별 경기 수를 홈/어웨이로 구분하여 정리한 것이다. 2013-2014시즌부터 2019-2020시즌까지 모든 시즌에 참가한 팀은 전체 33개 팀 중 11개 팀이었으며, 나머지

22개 팀은 적어도 1개 시즌에 참여하지 않았다.

표 1. 연구의 대상이 된 시즌별 경기수와 참가팀 수

연도	경기수	팀수
2013-2014	380	20
2014-2015	380	20
2015-2016	380	20
2016-2017	380	20
2017-2018	380	20
2018-2019	380	20
2019-2020	380	20
합계	2,660	-

표 2. 연구대상이 된 팀이 참가한 경기수 및 시즌수

팀 (n=33)	홈 경기수	어웨이 경기수	전체 경기수	참여 시즌수
arsenal	133	133	266	7
aston villa	76	76	152	4
bournemouth	95	95	190	5
brighton	76	76	152	4
burnley	114	114	228	6
cardiff	19	19	38	1
chelsea	133	133	266	7
crystal palace	133	133	266	7
everton	133	133	266	7
fulham	38	38	76	2
huddersfield	38	38	76	2
hull city	38	38	76	2
leeds utd	19	19	38	1
leicester city	133	133	266	7
liverpool	133	133	266	7
man city	133	133	266	7
man utd	133	133	266	7
middlesbrough	19	19	38	1
newcastle	114	114	228	6
norwich	38	38	76	2
QPR	19	19	38	1
sheffield utd.	38	38	76	2
southampton	133	133	266	7
stoke	76	76	152	4
sunderland	57	57	114	3
swansea	76	76	152	4
tottenham	133	133	266	7
watford	95	95	190	5
west bromwich	95	95	190	5
west ham	133	133	266	7
wolves hamton	57	57	114	3

2. 자료수집

축구 득점 차이 예측을 위한 머신러닝 기법을 비교하기 위하여 영국 EPL 공식 홈페이지(https://www.premierleague.com)에서 제공하는 공식기록을 Python 3.11.0을 사용하여 수집하였으며, Microsoft Excel을 이용하여 저장하였다.

3. 측정변수

본 연구를 위하여 영국 EPL 공식 홈페이지에서 수집된 측정변수는 〈표 3〉과 같다. 본 연구에서는 홈/어웨이 팀에 따른 차이(Goumas, 2013; Polland & Gomez, 2009)를 고려하기 위하여 총 26개의 독립변수를 홈팀과 어웨이팀으로 구분하여 수집하였으며, Goumas(2013)이 제시한 홈팀이 가진 이점으로 인한 결과의 차이를 반영하여 독립변수에 홈팀과 어웨이팀의 기록을 함께 고려하였다.

표 3. 측정변수

구분	변수명	단위
독립변수	Home Team Possession %	%
	Home Team Off Target Shots	빈도
	Home Team On Target Shots	빈도
	Home Team Total Shots	빈도
	Home Team Blocked Shots	빈도
	Home Team Corners	빈도
	Home Team Throw Ins	빈도
	Home Team Pass Success %	%
	Home Team Aerials Won	빈도
	Home Team Clearances	빈도
	Home Team Fouls	빈도
	Home Team Yellow Cards	빈도
	Home Team Red Cards	빈도
	Away Team Possession %	%
	Away Team Off Target Shots	빈도
	Away Team On Target Shots	빈도
	Away Team Total Shots	빈도
	Away Team Blocked Shots	빈도
	Away Team Corners	빈도
	Away Team Throw Ins	빈도
	Away Team Pass Success %	%
	Away Team Aerials Won	빈도
	Away Team Clearances	빈도
	Away Team Fouls	빈도
	Away Team Yellow Cards	빈도
	Away Team Red Cards	빈도
종속변수	Score Differences	

4. 예측을 위한 머신러닝 기법

1) 로지스틱 회귀분석 (Logistic Regression Analysis)

본 연구에서 사용한 머신러닝 기법 중 로지스틱 회귀분석 방법은 Cox(1958)가 제안한 확률 모델이며, 독립변수의 선형 결합을 이용하여 예측에 사용되는 통계기법이다. 로지스틱 회귀분석 방법에는 이진형(binary) 자료를 종속변수로 하는 이항 로지스틱 회귀분석(binominal logistic regression)과 종속변수의 수준이 세 개 이상일 때 사용하는 다항 로지스틱 회귀분석(multinominal logistic regression)이 있다. 본 연구에서는 홈팀이 승리할 때에는 1, 홈팀이 패배할 때에는 0, 비길 때에는 0.5로 설정하여 다항 로지스틱 회귀분석(Darroch & Ratcliff, 1972; Greene, 2012)을 실시하였다. 분류 K를 기준으로 자료가 각 분류에 속할 확률을 β_i, 입력되는 변수를 X_i, 출력되는 변수가 1이 될 확률을 Y_i라 할 때, 로지스틱 회귀 모형식은 〈수식 1〉과 같다.

$$P(Y_i = i) = \frac{e^{\beta_i X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k X_i}}$$

수식 1. 다항 로지스틱 회귀분석의 확률 계산식

2) 선형 판별 분석 (Linear Discriminant Analysis)

선형 판별 분석은 Fisher(1936)가 제안한 선형 판별 방법을 일반화시킨 것으로 자료의 수준을 구분하는 함수를 선형으로 조합해 도출하는 방법을 말한다. Z를 판별점수, β_0를 판별상수, β_1, β_2, …, β_p 를 판별계수, X_1, X_2, …, X_p를 판별변수라고 할 때, 선형 판별 분석에서 사용하는 판별함수는 〈수식 2〉와 같다.

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

수식 2. 선형 판별 분석의 판별함수 계산식

본 연구에서는 선형 판별 분석의 절차대로 최종적으로 판별점수의 집단 간 차이와 집단 내 차이의 비율을 최대화하는 판별함수를 도출한 후, 예측에 사용하였다.

3) 인공신경망 모델 (ANN: Artificial Neural Network)

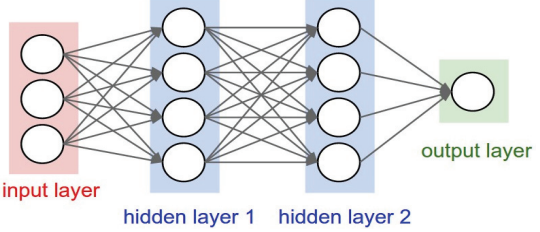
인공신경망 모델은 McCulloch와 Pitts(1943)의 연구에서 제안한 임계 논리 알고리즘(threshold logics algorithm)에 기반하여 개발되었다(최형준, 김주학, 2006). 인공신경망 모델에서 자료가 입력되는 층을 입력층(input layer), 결과가 출력되는 층을 출력층(output layer), 입력층과 출력층 사이에 존재하는 층을 은닉층(hidden layer), 각 뉴런 사이를 이어줄 때 발생하는 가중치를 연결가중치(linkage weight)라고 한다. 본 연구를 위해서 최형준과 김주학(2006)의 연구에서 사용한 인공신경망과 같은 신경망을 사용하였다. 본 연구에서 사용한 최종 출력값을 Y_m, j번째 은닉층과 출력층 간 연결가중치를 W_{mj}, j번째 입력층과 은닉층 간 연결가중치를 V_{ji}, i번째 입력층 변수를 X_i, i번째 입력층 연결가중치에 대한 편향(bias)을 β_i, j번째 은닉층 연결가중치에 대한 편향(bias)을 b_j라 할 때, 최종 출력값 계산식은 〈수식 3〉과 같다.

$$Y_m = f(\sum_{j=1}^{n} W_{mj} f(\sum_{i=1}^{m} V_{ji} X_i + \beta_i) + b_j$$

수식 3. 인공신경망의 최종 출력값 계산식

4) 딥러닝 모델 (Deep Learning Model)

딥러닝 모델은 인공신경망 모델에서 은닉층의 개수를 늘려서 설계하는 방법이다. 인공신경망 모델은 은닉층의 개수를 1개로 하여 설계하였지만, 본 연구에서 사용한 딥러닝 모델에서는 은닉층의 개수를 2개로 증가시킨 후, 결과를 도출하였다. 〈그림 1〉은 본



자료출처: https://cs231n.github.io/neural-networks-1/

그림 1. 본 연구에서 사용한 딥러닝 모델의 구상도

연구에서 사용한 딥러닝 모델의 구상도와 동일한 구조를 갖는 딥러닝 모델을 시각화한 것이다.

5) 서포트벡터머신 (Support Vector Machine)

서포트벡터머신이란 분류(classification)와 회귀분석(regression analysis)에서 사용하는 지도학습(supervised learning) 모델로서 데이터 집합을 바탕으로 하여 새로운 데이터가 속하는 카테고리를 판단할 때 사용되는 비확률적 이진 선형 분류 모델을 이용하여 계산되는 알고리즘을 말한다(Bertsimas et al., 2008). 서포트벡터머신의 장점은 커널이 존재할 때, 만일 x에 대한 벡터 값을 $\vec{x}$, y에 대한 벡터 값을 $\vec{y}$라고 할 때 커널 $K(\vec{x},\vec{y})$는 x에 대한 무한차원 $\phi(\vec{x})$과 y에 대한 무한차원 $\phi(\vec{y})$의 곱과 같게 된다. 즉, 서포트벡터머신의 장점은 우연에 의해 다른 공간에 있는 값에 대한 대응값을 분류할 수 있어 비선형적 관계에 있는 경우에도 분류의 정확도를 높일 수 있다는 점이다(William et al., 2007).

6) 나이브 베이즈 모델 (Naive Bayes Model)

나이브 베이즈 모델은 지도학습에 기반한 조건부 확률 모델로서, 결정 규칙을 조합하여 계산한다(Thabtah et al., 2019). 독립변수를 나타내는 $\vec{x}$에 대해 k개의 가능한 분류(C_k)를 나타낼 경우, 최대 확률을 갖는 클래스 k를 찾아내는 것은 각 독립변수의 값을 x_i, 독립변수의 수를 n, 최대확률을 갖는 클래스 k를 $\hat{y}$라고 할 때 〈수식 4〉와 같다(Thabtah et al., 2019).

$$\hat{y} = \underset{k \in 1,\ldots,k}{\arg max}\, p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

수식 4. 나이브 베이즈 모델에서 최대 확률을 갖는 클래스를 찾아내는 계산식

7) Extreme Gradient Boost (XGBoost)

XGBoost 모델은 기본적으로 의사결정나무(Decision Tree)에 기반하며 학습하며 잔차(gradient)를 토대로 이전에 사용된 모형을 보완하는 방식으로 학습하는 머신러닝 기법의 한 종류이다(Han et al., 2022). XGBoost 모델은 의사결정나무가 몇 가지가 되든지 모두 반영하여 계산할 수 있다는 장점을 지니고 있으며 각 트리로부터 받은 값들을 합쳐서 최종 예측값으로 결정된다(Chen & Guestrin, 2016). 주어진 n X m (n: 사례수, m: 측정 요소의 전체수) 자료를 $D=(x_i,y_i)(\mid D \mid = n, x_i \in R^m, y_i \in R)$라 할 때, 트리 앙상블 모델이 k개를 사용하여 트리의 Leaf를 고려한 예측값 계산식은 〈수식 5〉와 같다.

$$\hat{y_i} = \phi(x_i) = \sum_{k=1}^{k} f_k(x_i), f_k \in F,$$

수식 5. 트리의 수에 따른 앙상블 모델 예측값 계산식

XGBoost에서 실제값 y_i, 예측값 $\widehat{y_i}$, 정규화된 값 Ω, 가중치 w라 할 때, 손실함수 L은 〈수식 6〉과 같다. 〈수식 6〉에서 T는 leaf의 개수이며, γ와 λ는 정규화 파라미터이다. 만일 γ, λ 정규화 파라미터가 0이 되면, 잔차 트리 부스팅(Gradient Tree Boosting) 모델과 동일하기 때문에 XGBoost 모델이라고 할 수 없다(Chen & Guestrin, 2016).

$$L(\phi) = \sum_i l(\widehat{y_i}, y_i) + \sum_k \Omega(f_k),$$
$$where\, \Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2$$

수식 6. XGBoost 모델에서 사용하는 손실함수 L의 계산식

XGBoosting 모델에서 사용하는 학습의 계산식은 〈수식 7〉과 같다. 〈수식 7〉에서 t번째 iteration에 해당하는 손실함수는 총 사례수까지 학습을 반복하며, $\widehat{y_i}^{(t-1)}$는 t-1번째 iteration에서 도출된 예측값, $\Omega(f_t)$는 정규화 함수, Constant는 t-1번째 iteration까지의 정규화 함수의 합이다.

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t) + Constant$$

수식 7. XGBoosting 모델의 학습에 사용되는 계산식

5. 자료처리

본 연구에서는 로지스틱 회귀, 선형 판별 분석, 인공신경망 모델, 딥러닝 모델, 서포트벡터머신, 나이브 베이즈 모델, XGBoost 모델간 비교를 위하여 아래와 같이 자료를 처리하였다.

첫째, 모든 모델은 같은 학습자료와 테스트 자료로 예측하였다. 총 2,660개 경기자료를 무작위로 배열한 후, 80%에 해당하는 2,128개 자료를 학습자료로 선정하였으며, 20%에 해당하는 532개 자료를 예측자료로 선정하였다.

둘째, 각 모델별 예측값을 계산한 후, 실제값과 예측값을 배열(matrix) 형태로 정리하였다.

셋째, 각 모델 간 비교를 위하여 Cohen's Kappa, 정확도(accuracy), 민감도(sensitivity), 특이도(specificity)를 계산하였다. 또한 정밀도(precision), 재현율(recall), F1 score를 계산하여 모델 간 비교하였다. F1 score는 정밀도와 재현율을 모두 고려한 조화평균(homonic mean)이며, 0에서 1사이의 값을 갖는다. Cohen's Kappa, 민감도, 특이도, 정밀도, 재현율, F1 score는 모두 값이 높을수록 좋게 평가할 수 있다(Chicco & Jurman, 2020). 본 연구를 위한 자료처리는 R 4.2.2와 RStudio 2022.07.2.버전을 사용하였다.

Ⅲ. 연구결과

1. 기술통계

본 연구에서 사용한 독립 변수별 기술통계는 〈표 4〉와 같다. 홈팀이 승리한 경기의 수는 1,186경기

표 4. 본 연구대상의 기술통계 분석 결과

변수명	패배(n=845)		비김(n=629)		승리(n=1,186)	
	Mean	SD	Mean	SD	Mean	SD
Home Team Possession %	48.68	12.47	51.96	11.72	52.86	12.42
Home Team Off Target Shots	5.17	2.69	5.63	2.77	5.39	2.73
Home Team On Target Shots	3.32	2.02	4.19	2.19	5.89	2.66
Home Team Total Shots	12.01	5.22	13.86	5.68	15.17	5.61
Home Team Blocked Shots	3.54	2.4	4.04	2.75	3.9	2.53
Home Team Corners	5.47	3.08	6.03	3.28	5.86	3.05
Home Team Throw Ins	23.12	6.42	23.22	6.26	21.33	6.16
Home Team Pass Success %	77.01	7.08	77.74	7.04	79.36	8
Home Team Aerials Won	18.12	7.65	19.05	7.31	18.04	7.31
Home Team Clearances	19.65	9.04	22.67	10.21	23.48	10.9
Home Team Fouls	10.67	3.31	10.96	3.46	10.46	3.49
Home Team Yellow Cards	1.67	1.23	1.73	1.22	1.42	1.2
Home Team Red Cards	0.07	0.27	0.03	0.17	0.02	0.13
Away Team Possession %	51.32	12.47	48.04	11.72	47.14	12.42
Away Team Off Target Shots	4.52	2.43	4.45	2.35	4.23	2.35
Away Team On Target Shots	5.22	2.3	3.64	2.07	3.07	1.91
Away Team Total Shots	13.06	4.85	11.16	4.74	10.11	4.47
Away Team Blocked Shots	3.32	2.29	3.08	2.28	2.83	2.08
Away Team Corners	4.82	2.56	4.67	2.69	4.64	2.78
Away Team Throw Ins	21.04	5.78	21.49	6.15	21.71	6.64
Away Team Pass Success %	78.27	8.23	75.84	7.54	76.64	6.98
Away Team Aerials Won	18.28	7.28	19.07	6.99	17.46	7.41
Away Team Clearances	28.02	12.48	27.97	11.41	23.67	9.88
Away Team Fouls	10.84	3.59	11.3	3.49	10.93	3.49
Away Team Yellow Cards	1.69	1.27	1.88	1.36	1.74	1.24
Away Team Red Cards	0.02	0.13	0.04	0.2	0.06	0.25

(44.59%), 패배한 경기의 수는 845경기 (31.77%), 비긴 경기의 수는 629경기 (23.65%)로 나타났다. Possession %의 경우, 홈팀이 승리할 때는 어웨이 팀의 기록 평균에 비해 높았으며, 패배하거나 비긴 경기에서는 어웨이 팀의 기록 평균보다 낮게 나타났다. 또한, On Target Shots의 경우, 홈팀이 승리할 때는 어웨이 팀의 기록 평균보다 높았으며, 패배할 때는 어웨이 팀의 기록 평균보다 낮게 나타났다. Pass Success %의 경우, 홈팀이 승리할 때는 어웨이 팀의 기록 평균보다 높았으며, 패배할 때는 어웨이 팀의 기록 평균보다 낮게 나타났다.

2. 머신러닝 모델간 비교

본 연구에서는 실제 홈팀의 경기 결과가 승리, 패배, 비김인지를 토대로 머신러닝 모델을 학습시켰으며, 예측값과 비교하였다. 〈표 5〉부터 〈표 11〉까지는 실제값과 예측값 간 비교 내용을 교차표로 정리한 것이다.

표 5. 로지스틱 회귀분석 비교 결과

		예측값			
		패배	비김	승리	소계
실제값	패배	87	49	87	223
	비김	22	13	44	79
	승리	69	64	97	230
	소계	178	126	228	532

표 6. 선형 판별 분석 비교 결과

		예측값			
		패배	비김	승리	소계
실제값	패배	132	17	29	178
	비김	33	39	54	126
	승리	19	21	188	228
	소계	184	77	271	532

표 7. 인공신경망 모델 비교 결과

		예측값			
		패배	비김	승리	소계
실제값	패배	112	38	28	178
	비김	28	29	69	126
	승리	16	22	190	228
	소계	156	89	287	532

표 8. 딥러닝 모델 비교 결과

		예측값			
		패배	비김	승리	소계
실제값	패배	102	51	25	178
	비김	28	41	57	126
	승리	17	35	176	228
	소계	147	127	258	532

표 9. 서포트벡터머신 비교 결과

		예측값			
		패배	비김	승리	소계
실제값	패배	101	15	6	122
	비김	59	71	54	184
	승리	18	40	168	226
	소계	178	126	228	532

표 10. 나이브 베이즈 모델 비교 결과

		예측값			
		패배	비김	승리	소계
실제값	패배	109	40	29	178
	비김	43	39	44	126
	승리	40	31	157	228
	소계	192	110	230	532

표 11. XGBoost 모델 비교 결과

		예측값			
		패배	비김	승리	소계
실제값	패배	185	29	14	228
	비김	37	108	33	178
	승리	63	38	25	126
	소계	285	175	72	532

머신러닝 기법별 실제값과 예측값의 비교에서는 기법에 따라서 패배, 비김, 승리 중 예측률이 높은 기법들이 대부분으로 나타났다. 하지만 이러한 단순 교차 내용은 기법 간 비교가 어렵다. 따라서, 기법 간 평가 지표의 결과를 이용하여 비교하였다.

〈표 12〉는 본 연구의 머신러닝 기법 간 평가를 위해 선정한 지표 값을 정리한 표이다. Cohen's Kappa는 실제값과 예측값의 일치도를 살펴본 지표로서, 선형 판별 분석이 0.49로 가장 높았고, 로지스틱 회귀분석이 0.02로 가장 낮았다. Landis와 Gary(1977)가 제시한 기준표에 의해 해석하면, 로지스틱 회귀분석은 거의 없는 일치도를 나타냈다. 딥러닝 모델, 나이브 베이즈 모델, XGBoost는 어느 정도

표 12. 머신러닝 기법간 평가 지표 비교

머신러닝 기법	Cohen's Kappa	정확도	민감도	특이도	정밀도	재현율	F1 score
로지스틱 회귀	0.02	0.51	0.49	0.67	0.24	0.80	0.37
선형 판별 분석	0.49	0.74	0.72	0.84	0.39	0.91	0.54
인공신경망 모델	0.40	0.69	0.72	0.82	0.35	0.88	0.50
딥러닝 모델	0.38	0.68	0.69	0.8	0.35	0.86	0.50
서포트벡터머신	0.45	0.72	0.57	0.82	0.38	0.94	0.54
나이브 베이즈 모델	0.34	0.66	0.57	0.79	0.34	0.87	0.49
XGBoost	0.36	0.67	0.65	0.8	0.34	0.94	0.50

일치도를 나타냈고, 인공신경망과 서포트벡터머신은 적당한 일치도를 보였다. 정확도는 선형 판별 분석이 가장 높은 정확도(0.74)를 보였으며, 로지스틱 회귀분석이 가장 낮은 정확도(0.51)를 보였다. 민감도는 선형 판별 분석(0.72)과 인공신경망 모델(0.72)이 가장 높았으며, 로지스틱 회귀분석(0.49)이 가장 낮게 나타났다. 특이도도 선형 판별 분석(0.84)이 가장 높게 나타났으며, 로지스틱 회귀분석(0.67)이 가장 낮았다. 정밀도는 선형 판별 분석(0.39)이 가장 높았으며, 로지스틱 회귀분석(0.24)이 가장 낮았다. 재현율은 서포트벡터머신(0.94)과 XGBoost 모델(0.94)이 가장 높았으며, 로지스틱 회귀분석(0.80)이 가장 낮았다. F1 score는 선형 판별 분석(0.54)과 서포트벡터머신(0.54)이 가장 높게 나타났으며, 로지스틱 회귀분석(0.37)가 가장 낮은 것으로 나타났다.

Ⅳ. 논의

본 연구에서는 축구의 경기 결과 예측을 위한 머신러닝 기법의 비교를 위하여 영국 EPL 2013-2014 시즌부터 2019-2020 시즌까지 진행되었던 2,660 경기를 대상으로 홈팀 기준 승리, 패배, 비김 결과를 7가지 머신러닝 기법이 어떻게 예측하는지를 분석하였다. 최형준과 이윤수(2019)는 축구 경기 결과에 대한 예측을 위하여 자기구성지도(self-organizing map)을 사용하였으며, 독립변수의 모든 특성을 고려하여 예측하기 위해서는 독립변수의 차원 축소가 필요하다고 하였다. 하지만, 최형준과 김주학(2006)의 연구에 따르면 인공신경망 모델의 경우에는 입력층의 뉴런 개수가 몇 개가 되든지 간에 출력층에 투영되는 뉴런 개수가 2개인지 2개 이상인 것이 중요하다고 하였고, Greene(2012)의 연구에서도 다항 로지스틱 회귀분석을 위해서 차원 축소할 경우, 자료의 편향(bias)이 증가하거나 원하지 않는 결과를 초래할 가능성도 있다고 제언하였다.

본 연구에서도 머신러닝 기법에 자료가 입력되는 과정 중 독립변수의 차원 축소를 위해 자료를 왜곡하거나 변환하지는 않았다. 다만, XGBoost 모델의 경우에는 여러 겹의 트리 구조를 합쳐서 계산하는 방식을 따르고 있으므로 계산되는 과정에서 자료의 일반화는 매우 중요하게 작용하였다. 축구의 경기는 선수의 구성, 홈팀/어웨이팀 여부, 날씨, 경기장 잔디 상태와 같이 경기력과 관련이 있지만, 경기내용으로 파악할 수 없는 외재적 요인들로 인하여 결과가 다르게 나타날 수 있다. 만일 XGBoost 모델을 연도가 다른 상황에서 연도 간 비교를 위해 예측 모델을 설계한다든지 독립되는 사건을 대상 기준으로 종속변인을 공유한 상황에서 예측 모델을 설계할 경우, 앞서 제시한 독립변수의 본래 특성의 변화를 해결할 수 있는 방법론적인 대안이 될 수도 있다.

실제로 농구 경기 결과에 대한 예측을 위해 머신러닝 기법간 예측 성능에 관해서 연구한 최근 연구들(예원진, 이성노, 2022; 류지호, 한진욱, 김민수, 2018)에서 밝힌 바와 같이, 자료의 학습이 없거나 학습량이 적을 경우, 독립변수의 본래 특성 변화를 반영하지 못하여 예측적중률이 떨어지는 양상을 보인다. 본 연구에서 나타난 로지스틱 회귀분석 결과를 살펴보게 되

면, 민감도, 특이도, 정밀도, 재현율, F1 score 모두 다른 모델에 비해서 낮은 결과를 보였다. 본 연구에서는 다른 머신러닝 기법과의 예측 성능 비교를 위하여 같은 독립변수를 선정하고 비교하였기 때문에 로지스틱 회귀분석에서 제시되는 독립변수가 종속변수에 얼마나 영향을 주는지에 대한 고려가 적었다. 따라서, 독립변수의 선정과정에서 종속변수와의 상관(correlation)이 충분히 고려된다면, 로지스틱 회귀분석의 예측 적중률을 높이는 데 도움을 줄 것이다.

V. 결론

본 연구를 통해 다음과 같은 결론을 도출하였다.

첫째, 축구 경기 결과를 예측하는 데 가장 뛰어난 예측 성능을 나타낸 것은 선형 판별 분석이었으며, 서포트벡터머신과 XGBoost 모델의 경우에도 높은 F1 score를 나타냈다.

둘째, 로지스틱 회귀분석의 경우, 독립변수의 조합에 따라서 종속변수와의 상관(correlation)에 영향을 받는 만큼, 독립변수의 선정에 중점을 두어야 예측 성능을 높일 수 있을 것이다.

향후 스포츠 경기 결과 예측 관련 연구나 기타 스포츠 데이터의 예측 관련 연구를 위해서 자료의 특성, 예측 목적, 독립변수 간 관계를 고려한다면 예측 적중률을 높이고 스포츠의 본질을 살펴볼 수 있게 될 것이라 사료된다.

참고문헌

김종원, 최형준. (2021). K리그 축구 경기에서 나타난 상황 요인 별 주요 공격 분석인자의 상대성 분석. **한국체육학회지, 60**(1), 575-583. http://10.23949/kjpe.2021.1.60.1.41

김주학, 최형준. (2015). 자료범주에 따른 경기기록 승패의 재해석. **한국체육측정평가학회지, 17**(1), 1-12. http://10.21797/ksme.2015.17.1.001

류지호, 한진욱, 김민수. (2018). 포제션 개념을 적용한 한국남자프로농구 승률 예측 분석. **체육과학연구, 29**(1), 129-137. http://10.24985/kjss.2018.29.1.129

예원진, 이성노. (2022). 2022 FIBA 남자농구 아시안컵 경기결과를 활용한 머신러닝 분류 모형의 예측 성능 비교. **한국체육측정평가학회지, 24**(3), 53-69. http://10.21797/ksme.2022.24.3.005

이범수, 최형준. (2015). 축구 월드컵 대표팀의 선수구성에 따른 참가국 경기분석. **한국체육측정평가학회지, 17**(2), 13-23. http://10.21797/ksme.2015.17.2.002

이재현, 이수원. (2020). 앙상블 기법을 통한 잉글리시 프리미어 리그 경기결과 예측. **한국정보처리학회, 9**(5), 161-168. http://10.3745/KTSDE.2020.9.5.161

이해용. (2012). 포아송분포를 이용한 축구경기의 승률 예측에 관한 연구. **한국자료분석학회, 14**(1), 499-507. UCI: G704-000930.2012.14.1.019

최형준. (2020). 국내 스포츠 빅데이터 분석 연구의 현황. **한국체육측정평가학회지, 22**(2), 63-69. http://10.21797/ksme.2020.22.2.006

최형준, 김주학. (2006). 인공 신경망(Artificial Neural Network)을 이용한 2005년도 영국 윔블던 테니스 대회의 경기결과 예측에 관한 연구. **한국체육학회지, 45**(3), 459-467. UCI: G704-000541.2006.45.3.045

최형준, 엄한주. (2020). 스포츠경기분석의 이슈와 전망. **한국체육측정평가학회지, 22**(3), 105-113. http://10.21797/ksme.2020.22.3.009

최형준, 이윤수. (2019). 축구 월드컵대회의 경기기록 기반 경기결과 예측. **한국체육과학회지, 28**(1), 1317-1325. http://10.35159/kjss.2019.02.28.1.1317

홍종선, 정민섭, 이재형. (2010). 2010 남아공 월드컵 축구 예측모형 분석. **한국데이터정보과학회지, 21**(6), 1137-1146. UCI: G704-000605.2010.21.6.002

Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2018). From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research, 18*, 1-39. https://www.jmlr.org/papers/volume18/17-073/17-073.pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/29396

72.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient(MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21* (1), 1-13. https://doi.org/10.1186/s12864-019-6413-7

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B, 20*, 215-242. https://www.jstor.org/stable/2983890

Darroch, J. N., & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics, 43*(5), 1470-1480. http://10.1214/aoms/1177692379

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics, 7*, 179-188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Goumas, C. (2013). Modelling home advantage in sport: A new approach. *International Journal of Performance Analysis of Sport (E), 13*(2), 428-439. https://doi.org/10.1080/24748668.2013.11868659

Greene, W. H. (2012). *Econometric Analysis (Seventh ed.)*. Pearson Education Limited.

Han, J.-S., Jung, D.-H., & Kim, S.-J. (2022). Predicting the OPS of KBO Batters through Big Data Analysis Using Machine Learning. *The Journal of Next-Generation Convergence Technology Association, 6*(1), 12-18. https://doi.org/10.33097/JNCTA.2022.06.01.12

Landis, J. R., & Gary, G. K. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5*(4), 115-133. https://doi.org/10.1007/BF02478259

Polland, R., & Gomez, M. (2009). Home advantage in football in South-West Europe: Long-term trends, regional variation, and team differences. *European Journal of Sport Science, 9*(6), 341-352. https://doi.org/10.1080/17461390903009133

Shi, Y., Yu, P. S., & Tian, Y. (2014). Explore New Field of Data Science Under Big Data Era: Preface for ICDS 2014. *Procedia Computer Science, 30*, 1-3. http://10.1016/j.procs.2014.05.374

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA Game Result Prediction Using Feature Analysis and Machine Learning. *Annals of Data Science, 6*(1), 103-116. https://doi.org/10.1007/s40745-018-00189-x

William, H. ., Teukolsky, S. A. ., Vetterling, W. T. ., & Flannery, B. P. (2007). Section 16.5. Support Vector Machines. In Numerical Recipes: *The Art of Sciencefic Computing (3rd Ed.)*. Cambridge University Press.

저자정보

최형준(Hyongjun CHOI)
단국대학교 사범대학 체육교육과 부교수
chj2812@dankook.ac.kr

논문투고일 2022년 12월 06일
심사완료일 2022년 12월 21일
게재확정일 2022년 12월 26일

Abstract

The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science. 2022, 24(4), 81-91

Comparison of Machine Learning Methods for a Prediction of Match Outcomes in Soccer

Hyongjun CHOI *Dankook Univ.*

The purpose of this study was to compare the machine learning techniques used to predict match outcomes in soccer that can be applied to soccer goal difference prediction by comparing predicted hit rates. The subjects of this study were matches held in the English Premier League (n=2,660) from the 2013-2014 season to the 2019-2020 season, and the variables used for prediction were selected as a total of 26 independent variables and 1 dependent variable based on the home team. did A total of seven machine learning techniques were used for this study, and logistic regression analysis, linear discriminant analysis, artificial neural network, deep learning model, support vector machine, naive-Bayes model, and XGBoost were compared. The evaluation method of the model was compared with Cohen's Kappa, sensitivity, specificity, precision, recall, and F1 score. In conclusion, the best performed model among the machine learning models in this study, was Linear Discriminant Analysis. Also, Support Vector Machine and XGBoost model performed well with higher F1 score for a prediction of game outcomes in soccer.

Keywords: prediction of match outcomes, sport analytics, comparison of predictive performance index