The Korean Journal of Applied Statistics
2024, Vol. 37, No. 3, 323–338

DOI: https://doi.org/10.5351/KJAS.2024.37.3.323
Print ISSN1225-066X / Online ISSN 2383-5818

Matching prediction on Korean professional volleyball league

Heesook Kim[a], Nakyung Lee[a], Jiyoon Lee[a], Jongwoo Song[1,a]

[a]Department of Statistics, Ewha Womans University

Abstract

This study analyzes the Korean professional volleyball league and predict match outcomes using popular machine learning classification methods. Match data from the 2012/2013 to 2022/2023 seasons for both male and female leagues were collected, including match details. Two different data structures were applied to the models: Separating matches results into two teams and performance differentials between the home and away teams. These two data structures were applied to construct a total of four predictive models, encompassing both male and female leagues. As specific variable values used in the models are unavailable before the end of matches, the results of the most recent 3 to 4 matches, up until just before today's match, were preprocessed and utilized as variables. Logistc Regrssion, Decision Tree, Bagging, Random Forest, Xgboost, Adaboost, and Light GBM, were employed for classification, and the model employing Random Forest showed the highest predictive performance. The results indicated that while significant variables varied by gender and data structure, set success rate, blocking points scored, and the number of faults were consistently crucial. Notably, our win-loss prediction model's distinctiveness lies in its ability to provide pre-match forecasts rather than post-event predictions.

Keywords: volleyball, match prediction, machine learning, classification

1. 서론

배구는 각각 6명으로 구성된 두 팀이 경기장 중앙에 네트가 설치된 코트에서 네트 너머로 공을 쳐 넘기며 겨루는 구기 경기종목 중 하나이다. 5세트 중 3세트를 먼저 승리하는 팀이 경기에서 승리한다. 세트 승리는 2점 이상의 점수차로 25점에 도달하는 팀이 가져간다. 이 때 먼저 25점에 도달하더라도 동점이거나 1점 차이인 경우에는 점수차가 2점이 될 때까지 세트를 계속 진행하며, 5세트의 경우 15점에 먼저 도달하는 팀이 경기에서 승리하게 된다.

한국의 배구 프로 리그 명칭은 'V-리그'이다. 2005년 정식 출범하여 올해 3월에 18번째 시즌을 마무리하였다. 2022/2023 시즌 기준 여자부와 남자부 각각 7개 팀이 참가하고 있다. V-리그는 8월경에 프리시즌 개념인 KOVO컵을 시작으로 하여, 10월 중순 정규 시즌이 개막한다. KOVO컵이 약 1 ~ 2주 정도의 짧은 기간 동안 진행되는 반면, 정규 시즌은 약 6개월 동안 진행된다. 각 팀은 6개의 라운드 동안 36번의 경기를 치르며, 정규 시즌 승점 1위 팀이 정규 시즌 우승 팀이다. 정규 시즌 결과 상위 승점 구단은 포스트 시즌에 진출할 자격을 얻는다. 기본적으로는 상위 3개 팀만 포스트 시즌에 진출할 수 있지만, 3위 팀과 4위 팀의 승점 차이가 3점 이내이면 상위 4개 팀까지 포스트 시즌에 진출하여 경기를 치른다. 포스트 시즌은 플레이오프-챔피언 결정전 순으로 모두 토너먼트 형식이다. 최종 우승 팀은 챔피언 타이틀을 얻게 된다.

[1]Corresponding author: Department of Statistics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: josong@ewha.ac.kr

Published 30 June 2024 / journal homepage: http://kjas.or.kr

스포츠 현장에서의 승리 예측은 예로부터 많은 관심을 받아온 주제로, 전력에 대한 파악 및 평가, 새로운 작전을 구성하여 궁극적으로 팀의 승리를 꾀한다 (Huh와 Yoon, 2023). 배구 경기의 양적 분석과 평가는 선수 자신의 기술 향상을 위한 기초적 자료를 제공할 수 있으며, 팀을 훈련, 관리하는 코치들에게 빼놓을 수 없는 중요한 과정이라 할 수 있을 뿐만 아니라, 관중과 보도진의 측면에서 볼 때, 신속하고 객관적인 보도를 위해 요구되는 중요한 요소로 인식됨에 따라 경기력과 관련된 다양한 연구들이 진행되었다 (Kwon 등, 1998; Baacke, 1982; Hughes와 Frank, 1997; Eom와 Shutz, 1992). 그동안 배구에 관한 선행연구들을 살펴보면 공격유형에 관한 연구 (Shin, 2017; Chun과 Kim, 2011; Kim, 2007)와 서브에 관한 연구 (Cho, 2017; Kim 등, 2011; Kim, 2009), 특정 포지션의 역할에 대한 연구 (Hong, 2011; Jo, 1999) 등이 있다. 그동안의 선행연구들이 한국 배구 발전에 많은 기여를 한 것은 사실이지만 대부분 공격유형과 공격성공률에 관련된 연구가 주를 이루었다. 이는 경기가 종료된 이후 나오는 해당 경기에 대한 데이터가 필요하므로 실제 사전에 경기 결과 예측을 하는 데는 부족한 실정이다.

더불어 스포츠 경기는 직접 참여하거나 관람하는 것으로 끝나는 것이 아니라 승부를 예측하여 베팅을 하는 것까지도 스포츠를 즐기는 하나의 방법으로 여겨지고 있다. 승부 예측을 바탕으로 승리가 예상가는 팀에 배팅을 하는 체육진흥투표권(토토, 프로토)은 우리나라의 경우 국가가 운영하며, 이로부터 얻은 수익은 국민체육금으로 편입되어 대한민국 스포츠 산업의 활성화에 이바지를 하고 있다. 이러한 승부 예측에 있어서는 대표적으로 통계적 모형과 머신러닝을 이용하여 진행하는 연구가 가장 보편적이다. 기존에 진행되었던 연구들을 보면 로지스틱 회귀 분석 (McCullagh와 Nelder, 1989; Hastie와 Pregibon, 1992), 의사결정나무분석 (Breiman 등, 1984) Bagging (Breiman, 1996), Random Forest (Breiman, 2001), LSTM (Ke 등, 2017) 등 다양한 통계적 모형과 머신러닝을 사용하여 추정한다. 다만, 이전까지는 스포츠 경기 승부 예측은 경기가 끝난 후에 측정되는 데이터를 통하여 사후 추정 가능한 분석이었다면, 본 연구에서는 경기 시작 전까지 관측 가능한 데이터만을 이용한 예측 모형을 만드는데 의의가 있다. 더불어 본 연구에서는 단순한 스포츠 경기 결과의 예측에만 주목하지 않을 뿐 아니라, 한국 프로 배구 경기의 승부 예측을 남자 배구와 여자 배구로 나누어 진행하여, 남자/여자 프로 배구 경기에 따른 중요 변수의 변화에 대해 알아보고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 분석 데이터의 수집 방법 및 주요 변수들에 대해 설명하고 3장에서는 다양한 데이터마이닝 기법을 활용하여 적합한 모형 및 그 결과에 대해 설명하고 4장에서 본 연구의 결과를 요약한다.

2. 분석자료

2.1. 데이터 수집 과정

본 연구에서 활용할 데이터를 수집하기 위해 Python을 이용하여 웹크롤링(web crawling)을 진행하였다. 분석의 핵심이 되는 경기 기록 데이터를 구축하기 위해 KOVO 한국 배구 연맹 에서 제공하는 V-리그 경기 상세 결과를 수집하였다. 데이터는 V-리그 남자부와 여자부, 2012–2013 시즌부터 2022–2023시즌까지 총 11개 시즌을 대상으로 하며 남자부는 7개의 팀으로 1,377경기, 여자부는 2021–2022 시즌부터 참가한 페퍼저축은행을 제외한 6개의 팀으로 1,026경기로 구성된다. 수집한 데이터는 다음과 같이 2가지 경우로 정의하여 예측에 이용한다. 첫 번째 경우는 각 경기에 참가한 팀을 기준으로 정리한 것으로, 최종적으로 남자부 2,754개와 여자부 2,052개의 경기 내 팀 상세 결과를 이용하였다. 두 번째 경우는 홈경기장에서 경기한 팀을 기준으로 수집한 변수에 대해 홈팀과 원정팀의 차이(홈팀-원정팀)를 계산하여 이용하였다. 이 경우 최종적으로 남자부 1,377개와 여자부 1,026개의 결과를 이용하였다. 2012/2013 시즌 이전 경기들은 프로배구경기 승부조작 사건 이 확정됨 (Ji, 2014)에 따라 승패 결정 요인 분석 및 예측에 적절하지 않을 것으로 판단하여 분석에서 제외하였다.

Table 1: Description of variables

Variable	Description	Type
BlockingSc	한 세트당 평균 블로킹 득점	Numerical
ServeSc	한 세트당 평균 서브 득점	
MisSc	한 세트당 평균 범실	
DigSn	한 세트당 평균 디그 성공의 수	
ReceiveAn	한 세트당 평균 리시브 정확 단계의 수	
SetSn	한 세트당 평균 세트 성공의 수	
Attack_scrate	경기 종료시 공격 성공률(%)	
Receive_scrate	경기 종료시 리시브 효율(%) 리시브 효율 = (리시브 정확 – 리시브 실패)/전체 리시브 횟수	
Homeground	팀 소재지의 경기장 여부(홈:1, 원정:0)	Categorical
Derived variable		
Ssrate	경기 종료시 Top1과 Top2 선수의 공격점유율 대비 공격성공률의 평균	Numerical
Winloss	오늘 경기를 제외한 시즌 경기에 대해 누적 연승 및 연패	
Rival	시즌 내 경기 상대와의 승률	
Before_2	시즌 내 직전 두 경기 결과	
Response variable		
Result	경기 승패(승:1, 패:0)	Categorical

2.2. 변수 설명

본 연구의 목적은 배구 프로 경기 승패 결과에 영향을 미칠 것이라 예상되는 변수들을 이용하여 배구 승패의 결과를 예측하는 것으로, 남자 프로 배구, 여자 프로 배구, 전체 프로 배구로 구분하여 예측 모형을 고려하였다. 모형에 사용된 반응 변수는 경기 진행 후 결과로 하나의 경기에 참가한 팀은 1(승)또는 0(패)의 값을 갖는다. 설명 변수는 구축한 경기 기록 데이터를 업데이트 하여 사용하였다. 시즌이 시작하면 매주 한 팀은 2번의 경기를 진행한다. 따라서, 팀의 경기 기록 데이터를 경기 시작전까지 관측 가능한 이전 3경기 혹은 4경기의 경기 상세 결과에 대해 최근 경기일수록 높은 가중치를 주어 평균값으로 업데이트하였다. 분석에 사용된 설명 변수는 Table 1와 같다. 먼저 배구는 5세트 중 3세트를 먼저 획득하는 팀이 승리하는 스포츠이다. 25점을 먼저 득점하는 팀이 세트를 획득하게 되며, 듀스가 존재하여 2점 이상의 차이가 있어야 승리한다. 따라서, 진행한 세트에 따라 득점 값은 크게 달라진다. 이를 고려하여 경기 종료 시 측정된 각 득점을 진행한 세트 횟수로 나누어 세트당 획득한 평균 득점으로 변환하여 사용한다. 배구에서 득점은 공격 득점, 블로킹 득점, 서브득점, 상대 범실로 인한 득점으로 이루어지며 내용은 다음과 같다.

- 공격 득점

  공격은 서브와 블로킹을 제외하고 상대편 코트로 볼을 넘기는 행위로, 스파이크와 리시브 및 세트 동작에서 상대 코트로 넘기는 모든 것이 공격으로 간주된다. 공격으로 승부가 결정되면 공격 득점으로 기록되며, 이를 공격 성공이라고 한다.

- 블로킹 득점

  블로킹은 상대 스파이크를 방어하기 위해 전위 선수가 네트 위로 점프하여 두 손으로 공을 방어하는 행위로, 블로킹으로 승부가 결정되면 블로킹 득점으로 기록된다.

- 서브 득점

  서브는 엔드라인 후방 서브존에서 타구를 하는 행위로, 서브로 승부가 결정되면 서브 득점으로 기록된다.

- 상대 범실로 인한 득점

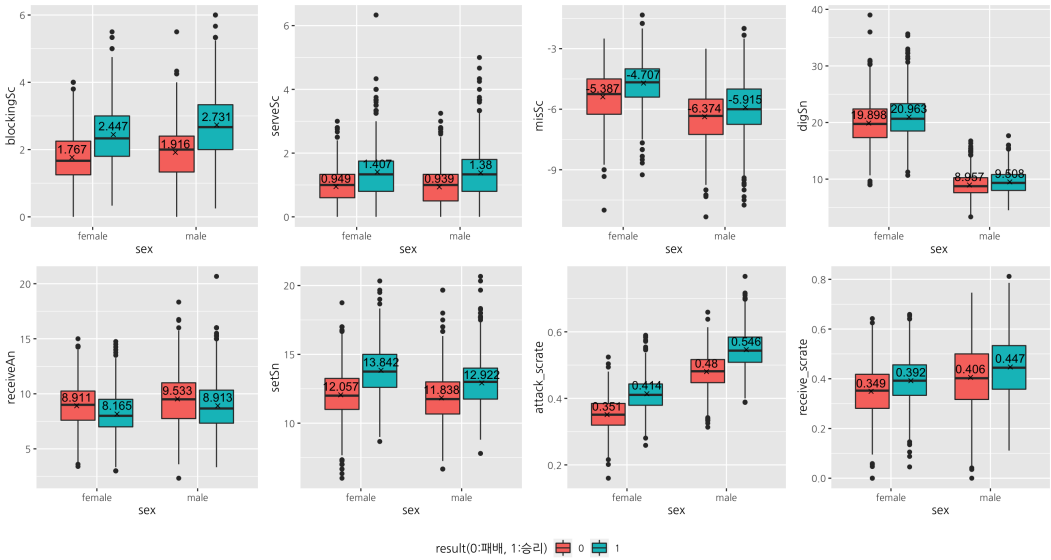

Figure 1: *Distribution of response and explanatory variables for men's and women's league.*

범실은 팀의 반칙 및 공격이 라인을 벗어나거나 서브가 네트에 걸리는 등의 행위로, 그 결과 상대 팀의 득점으로 기록된다. 따라서, 분석에서는 범실로 인해 상대 팀이 획득한 점수를 음수로 변환하여 우리 팀이 상대 팀에게 잃은 점수를 의미하는 변수로 사용하였다.

- 공격 성공률 공격 성공률은 전체 공격 시도 중에 공격 성공이 차지하는 비율이다.
- 공격 점유율 대비 성공률의 평균 공격 점유율은 팀의 공격 시도 중에 선수 개인의 공격 시도가 차지하는 비율이다. 따라서, 분석에서는 팀내 공격 점유율이 높은 순으로 Top1과 Top2선수의 공격 점유율 대비 성공률의 평균을 변수로 사용하였다.

또한 배구는 한사람의 연속 터치를 금지하며 3번 이내의 터치로 상대 팀으로 공을 넘겨야 한다. 이 과정에서 선수들간 연계가 중요하며 이는 공격의 기회로 이어질 수 있다. 따라서 배구에서는 디그 성공, 리시브 정확, 세트 성공을 기록하고 있으며 이는 다음과 같다.

- 디그 성공

  디그는 상대 공격에 대해 블로킹을 제외한 첫 번째 수비를 뜻한다. 따라서 디그 성공은 상대팀의 공격을 받아내 같은 팀 동료에게 연결한 경우를 의미한다.

- 리시브 정확

  리시브는 정확/리시브/실패 3단계로 구분된다. 이 중 정확 단계는 세터가 세 걸음 이내에 세트업 할 수 있는 리시브를 의미하며, 리시브가 정확할 수록 세터는 공을 자유롭게 다룰 수 있다.

- 세트 성공

  세트 성공은 같은 팀 A선수가 올린 공을 B선수가 공격했을 때 득점을 하게 되는 세트 업을 의미한다.

Table 2: Total matches and championship wins by men's teams

팀 명	총 경기 수	정규리그 경기 수	정규리그 승수	포스refer그 승수	최종 챔피언 우승	전체 경기 승률
대한항공	413	379	243	17	4	0.62
현대캐피탈	414	380	224	16	2	0.57
삼성화재	396	350	202	8	2	0.53
OK금융그룹	364	380	161	11	2	0.47
우리카드	391	380	176	4	0	0.46
KB손해보험	386	381	160	2	0	0.41
한국전력	390	380	150	3	0	0.39

Table 3: Total matches and championship wins by women's teams

팀 명	총 경기 수	정규리그 경기 수	정규리그 승수	포스트리그 승수	최종 챔피언 우승	전체 경기 승률
IBK기업은행	357	324	184	18	3	0.56
현대건설	337	323	174	6	1	0.53
GS칼텍스	342	323	172	12	2	0.53
한국도로공사	343	323	170	11	2	0.52
흥국생명	346	325	155	8	1	0.47
KGC인삼공사	327	322	115	1	0	0.35

2.3. 결측치 처리 방법

팀내 공격 점유율이 높은 Top1과 Top2선수의 공격 점유율 대비 성공률의 평균을 의미하는 파생변수(ssrate)를 생성하는 과정에서 결측값을 확인하였다. 점유율 및 성공률 값이 0 등의 값으로 기록된 경우 점유율 대비 성공률을 계산하는 과정에서 결과값으로 Inf 또는 NA값이 발생하였다. 본 연구에서는 이와 같이 발생한 결측값를 처리하기 위해 K-nearest neighbor imputation (KNN) 대치법을 사용하였다. KNN 대치법은 결측값과 가장 가까운 거리에 있는 k개의 이웃 개체들을 이용하여 결측값을 대체하는 방법이다. 따라서, 시즌 내 같은 팀 조합으로 진행된 경기가 4~8개라는 점을 고려하여 결측이 발생한 변수와 상관관계가 높은 3개의 변수를 이용하였고, 유클리드 거리를 사용하여 이들의 중앙값으로 결측값을 대체하였다. ssrate를 제외한 다른 변수에서는 결측값이 존재하지 않았다.

2.4. 탐색적 자료 분석(EDA)

예측 모형 설계에 앞서 데이터 분석 기법 중 탐색적 자료 분석(EDA) 기법을 적용하여 데이터 간 패턴을 확인하였다. 두 팀이 진행하는 하나의 경기는 각 팀으로 구분하여 2개의 행으로 재구성하였으므로, 남자 배구 데이터는 2,754개, 여자 배구는 2,052개의 행으로 이루어진다.

Figure 1은 V-리그 남자부와 여자부의 반응변수와 설명변수를 비교한 상자그림으로 대체로 남자 배구의 수치가 여자 배구 수치보다 높은 수치를 보이며, 성별에 따른 변수 분포 차이가 존재함을 알 수 있다. 디그 성공의 수(digSn)와 공격 성공률(attack_scrate)의 경우 평균값의 차이가 큰 것을 알 수 있으며, 리시브 효율(receive_scrate)의 경우 남자 배구에서 상대적으로 변동성이 큰 것을 알 수 있다. 특히, 디그 성공의 수와 공격 성공률을 보면 남자 배구가 여자 배구보다 디그 성공의 수는 작으나 공격 성공률은 높은 것을 볼 수 있다. 이는 남자 배구의 경우 공격이 득점으로 이어지는 경우가 많기 때문에 상대의 공격을 수비 후 팀에게 전달하는 디그 성공의 수가 적은 것으로 여자 배구와 다른 분포를 보인다. 또한, 리시브 정확의 수(receiveAn)와 리시브

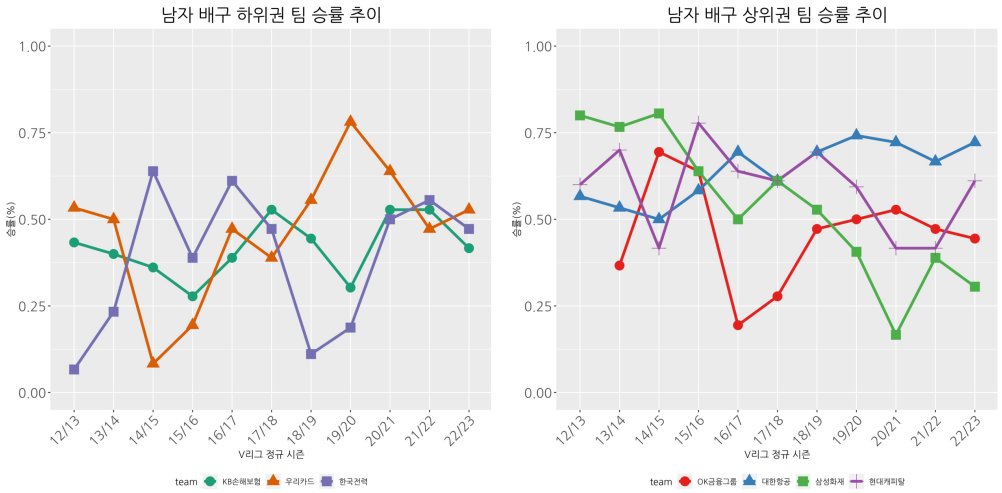

Figure 2: *Men's volleyball team win-loss trends by season.*

효율(receive_scrate)변수를 보면, 리시브 정확의 수는 남자와 여자 모두 패배한 경우 높은 값을 보이나, 리시브 효율은 승리한 경우 높은 값을 볼 수 있다. 즉, 리시브를 정확하게 하는 것 보다는 리시브 실패의 수를 줄이는 것이 승리 가능성을 높인다고 할 수 있다.

Tables 2–3에서는 11개의 시즌에 대해 참가한 팀의 총 진행 경기 수와 정규 리그, 포스트 리그 각 승리의 횟수 및 최종 챔피언 우승 횟수를 나타낸다. 남자 배구 팀 'OK금융그룹'의 경우 2013년 설립 배구단으로 12/13 시즌에 참여하지 않아 30개 경기수 차이가 있다. 또한, 코로나19로 인한 정규 리그 조기 종료로 남자부와 여자부 내 팀별 정규 리그 경기수가 다르며, 최종 챔피언 우승 횟수는 최종 챔피언 결승전이 취소된 남자부 19/20 시즌과 여자부 19/20, 21/22 시즌을 제외한 각각 10개, 9개 시즌에 대한 결과이다. 남자부의 경우 최종 챔피언 우승을 많이 한 팀은 대한항공으로 총 10개 시즌 중 4시즌(17/18, 20/21, 21/22, 22/23시즌)에서 챔피언 결정전 최종 우승을 하였으며, 여자부의 경우 IBK기업은행으로 총 9개 시즌 중 3시즌(12/13, 14/15, 16/17시즌)에서 챔피언 결정전 최종 우승을 하였다. 또한 남자부와 여자부 모두 정규리그에서 좋은 성적을 보이는 팀이 포스트 리그에서도 좋은 성적을 보이는 것을 알 수 있다. 한편, 전체 경기에 대한 각 팀의 승률을 보았을 때 대체로 값이 0.4 ~ 0.6 사이임을 알 수 있었다. 즉, 월등한 팀이 존재하여 계속 우승하는 경우는 없는 것으로, 두 팀의 경기 결과에 대한 승패를 예측하는 것의 어려움을 알 수 있다.

Table 2와 Table 3의 최종 챔피언 우승 횟수를 기준으로, 성적이 좋은 팀과 좋지 않은 팀을 비교하여 승률을 비교해보면 남자 배구는 Figure 2, 여자 배구는 Figure 3과 같다. 남자부의 경우 성적이 좋은 상위권 팀은 대한항공, 현대캐피탈, OK금융그룹, 삼성화재 총 4개의 팀이며, 여자부의 경우 IBK기업은행, 한국도로공사, GS칼텍스 총 3개의 팀으로 정의하였다. 성적이 좋지 않은 하위권 팀은 앞서 언급한 팀을 제외한 팀으로 남자부의 경우 한국전력, KB손해보험, 우리카드 3개 팀과 여자부의 경우 현대건설, 흥국생명, KGC인삼공사 3개 팀으로 정의하였다.

먼저 남자 배구에 대해 챔피언 우승으로 본 승률 추이를 Figure 2의 오른쪽 그래프를 통해 보면, 챔피언 최다 우승팀인 대한항공은 대체로 상승하는 추세를 볼 수 있다. 또한, 챔피언 우승 횟수 2회로 같은 수치를 가진 현대캐피탈, OK금융그룹은 시즌에 따른 승률 변동성이 크며, 삼성화재의 경우 대체로 감소하는 추세를 확인 할

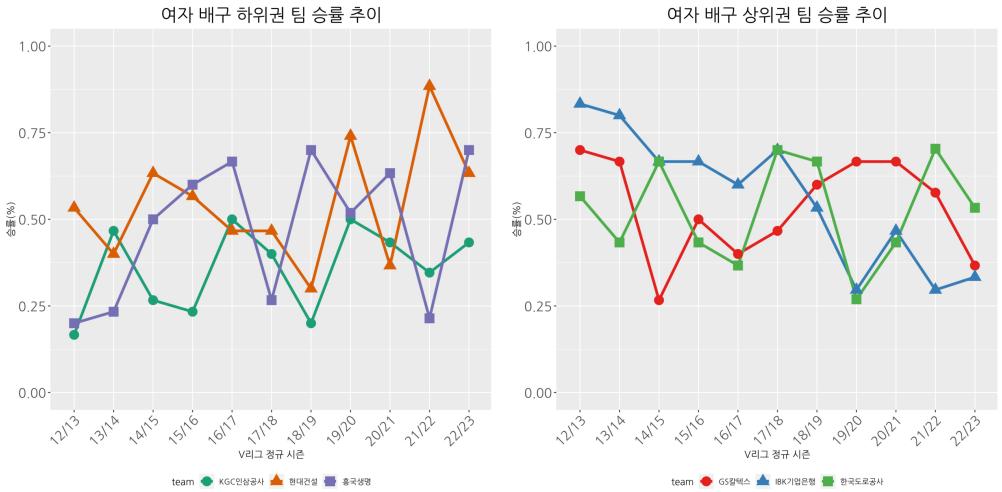

Figure 3: *Women's volleyball team win-loss trends by season.*

수 있다. 다음으로 챔피언 성적이 좋지 않은 남자 배구 팀 추이를 Figure 2 왼쪽 그래프를 통해 보면, 대부분의 팀 승률이 성적이 좋은 팀(우측 그래프)보다 낮은 값에 분포해 있으며, 변동성이 큰 것을 볼 수 있다.

여자 배구의 경우 Figure 3 오른쪽 그래프를 통해 챔피언 성적이 좋은 팀을 보면, IBK기업은행은 승률이 하강하는 추세를 확인할 수 있다. 또한 항상 눈에 띄게 높은 승률을 보이는 팀은 없으며, 시즌이 바뀜에 따라 팀의 승률 기복이 심한 것을 알 수 있다. Figure 3의 왼쪽 그래프를 통해 성적이 좋지 않은 팀의 분포를 보면, KGC인산공사를 제외한 흥국생명과 현대건설은 성적이 좋은 팀과 수치적으로 큰 차이를 보이지 않는 것을 알 수 있다. 결과적으로 남자 배구와 여자 배구, 챔피언 성적의 좋고 나쁨에 관계없이 승률은 시즌 기복이 심한 것을 알 수 있으며, 이는 과거 시즌 승패를 통해, 더 나아가 이전 시즌 챔피언 팀이라고 하더라도 다음 시즌의 경기 승패를 예상하기 쉽지 않음을 의미한다. 즉, 배구 승부 예측의 어려움을 시사한다.

3. 분석 결과

본 연구는 한국 여자 배구 프로 리그와 한국 남자 배구 프로 리그의 승패를 예측하기 위한 예측 모형을 설계하였다. 또한 수집한 데이터를 각 팀의 경기 결과를 기준으로 한 데이터와 각 경기 결과를 홈팀 기준으로 변형한 데이터로 구조화하여 분석을 다각화하였다. 구체적으로 사용한 모형은 대표적인 머신러닝 분류 기법인 배깅(Bagging), 랜덤포레스트(Random Forest), 익스트림 그래디언트 부스팅(Xgboost), 에이다부스트(Adaboost), 라이트 그래디언트 부스팅 모델(Light GBM)이며, 예측력 평가 지표로는 정분류율(Accuracy)를 사용하였다. 훈련데이터로는 2012/2013시즌부터 2020/2021시즌까지의 경기를, 시험데이터로는 2021/2022 시즌부터 2022/2023시즌까지의 경기를 사용하였다. 다섯 가지의 학습 결과 중 각각의 데이터 구조에서 가장 좋은 예측력을 보인 모형에 대한 변수 중요도 그림을 통해 어떤 변수가 경기 승리에 더 많이 기여하는지 파악하고, 부분 의존도 그림을 통해 각 변수가 경기 승리 확률과 어떤 관계를 가지는지 알아보았다.

한편, 여자 배구 프로 리그의 경우 2022/2023 시즌 기준 7개 구단이 리그에 참가하고 있으나, 페퍼저축은행 구단은 2021/2022 시즌에 신규 창단하여 경기 수가 다른 구단에 비해 매우 적으므로 이상치로 간주하고 데이

Table 4: Mean test error rate of each model using 5-fold cross validation (Model 1, men's volleyball)

Model	Cross-validation error
Logistic Regression	0.4024
Decision tree	0.4282
Bagging	0.4259
Random Forest	**0.4005**
Xgboost	0.4296
Adaboost	0.4184
Light GBM	0.4517

Table 5: Confusion matrix and accuracy (Model 1, men's volleyball)

	혼동행렬		정확도
	실제 패배	실제 승리	
예측 패배	158	120	0.572
예측 승리	106	144	

터에서 제외한 후 분석에 사용하였다. 남자 배구 프로 리그의 경우 2013/2014 시즌부터 OK금융그룹 구단이 신규 창단되어 리그에 합류하였다. 본 연구의 분석 대상이 2012/2013 시즌부터 2022/2023 시즌까지의 경기임을 감안했을 때 OK금융그룹의 경기 수가 다른 구단과 큰 차이를 보이지 않으므로 전체 데이터를 사용하였다.

3.1. 한국 프로 배구 리그의 팀별 승패 예측 모형 (Model 1)

Model 1은 경기 결과 데이터를 각 팀을 기준으로 정렬한 후 승부 예측을 시행한 것이다. 분석에 사용된 경기 수는 남자 배구 1377개, 여자 배구 1024개이므로 이를 각 팀을 기준으로 정렬하여 남자 배구 2754개, 여자 배구 2048개 경기 결과 데이터를 사용하였다.

3.1.1. 남자 프로 배구 팀별 승패 예측 모형(Model 1)

Table 4는 2012/2013 시즌부터 2020/2021 시즌까지의 남자 배구 경기 데이터를 이용하여 Bagging, Random Forest, Xgboost, Adaboost, Light GBM으로 시행한 5겹 교차검증 결과이다. 랜덤포레스트의 교차검증 오분류율이 0.4005로 가장 낮았으며, 최적 파라미터는 mtry = 1이다. 따라서 Model 1의 남자 배구 데이터에 대해서는 랜덤포레스트를 최종 모형으로 선택하였다.

Table 5는 Model 1의 남자 배구 승부 예측 최종 모형으로 선택한 랜덤포레스트를 사용하여 시험데이터의 승패를 예측한 결과이다. 예측 정확도는 0.572이며, 혼동행렬을 보았을 때 실제 경기 결과와 다르게 분류한 경우가 승리와 패배 여부에 따라 큰 차이를 보이지 않는 것으로 나타났다.

3.1.2. 여자 프로 배구 팀별 승패 예측 모형(Model 1)

Table 6는 2012/2013 시즌부터 2020/2021 시즌까지의 여자 배구 경기 데이터를 이용하여 Bagging, Random Forest, Xgboost, Adaboost, Light GBM으로 시행한 5겹 교차검증 결과이다. 여자 배구 경기에서는 Adaboost의 교차검증 오분류율이 0.4202로 가장 낮았으므로 Model 1의 여자 배구 데이터에 대해서는 Adaboost를 최종 모형으로 선택하였으며, 최적 파라미터는 maxdepth = 1, iter = 100, nu = 0.7이다.

Table 7는 Model 1의 여자 배구 승부 예측 최종 모형으로 선택한 Logistic Regressiond을 사용하여 시험데이터의 승패를 예측한 결과이며, 예측 정확도는 0.5791이다. 혼동행렬을 보았을 때, 남자 배구와는 달리 패배할

Table 6: Mean test error rate of each model using 5-fold cross validation (Model 1, women's volleyball)

Model	Cross-validation error
Logistic Regression	**0.4124**
Decision Tree	0.4502
Bagging	0.4420
Random Forest	0.4370
Xgboost	0.4429
Adaboost	0.4202
Light GBM	0.4502

Table 7: Confusion matrix and accuracy (Model 1, women's volleyball)

	혼동행렬		정확도
	실제 패배	실제 승리	
예측 패배	86	58	0.5791
예측 승리	91	119	

Table 8: Mean test error rate of each model using 5-fold cross validation (Model 2, men's volleyball)

Model	Cross-validation error
Logistic Regression	0.3952
Decision Tree	**0.3908**
Bagging	0.4642
Random Forest	0.4717
Xgboost	0.5774
Adaboost	0.4528
Light GBM	0.4830

것으로 예측한 경기가 실제로 패배한 경우가 승리할 것으로 예측된 경기에서 실제로 승리한 경우보다 두드러지게 높은 것으로 나타났다.

3.2. 한국 남녀 프로 배구 리그의 홈팀 기준 경기별 승패 예측 모형 (Model 2)

Model 2는 각 경기 데이터를 홈 팀 기준으로 원정 팀의 값을 빼거나 나누어 정렬한 후 승부 예측을 시행한 결과이다. Model 1과 다르게 홈팀과 원정팀을 나누어 전처리를 진행했으므로 홈/원정 변수는 제외하고 경기 남자 배구 1377개, 여자 배구 1024개 경기에 대한 결과 데이터를 사용하였다.

3.2.1. 남자 프로 배구 리그의 팀별 승패 예측 모형 (Model 2)

Table 8은 2012/2013 시즌부터 2020/2021 시즌까지의 남자 배구 경기 데이터를 이용하여 Bagging, Random Forest, Xgboost, Adaboost, Light GBM으로 시행한 5겹 교차검증 결과이다. Adaboost, Light GBM으로 시행한 5겹 교차검증 결과이다. 랜덤포레스트의 교차검증 오분류율이 가장 낮았으며, 최적 파라미터는 iter = 20, nu = 1 이다. 따라서 Model 2의 남자 배구 데이터에 대해서는 Adaboost를 최종 모형으로 선택하였다. Table 9는 Model 2의 남자 배구 승부 예측 최종 모형으로 선택한 Adaboost를 사용하여 시험데이터의 승패를 예측한 결과이며, 예측 정확도는 0.5358이다. 혼동행렬을 보았을 때, 승리할 것으로 예측한 경기가 실제 패배한 경우보다 패배한 것으로 예측한 경기가 실제 승리한 경우일 때 오분류율이 높다.

Table 9: Confusion matrix and accuracy (Model 2, men's volleyball)

	혼동행렬		정확도
	실제 패배	실제 승리	
예측 패배	60	70	0.5396
예측 승리	52	83	

Table 10: Mean test error rate of each model using 5-fold cross validation (Model 2, women's volleyball)

Model	Cross-validation error
Logistic Regrssion	0.3578
Decision Tree	0.3451
Bagging	0.4294
Random Forest	**0.3080**
Xgboost	0.5876
Adaboost	0.4520
Light GBM	0.4576

Table 11: Confusion matrix and accuracy (Model 2, women's volleyball)

	혼동행렬		정확도
	실제 패배	실제 승리	
예측 패배	45	34	0.6497
예측 승리	28	70	

3.2.2. 여자 프로 배구 리그의 팀별 승패 예측 모형 (Model 2)

Table 10는 2012/2013 시즌부터 2020/2021 시즌까지의 여자 배구 경기 데이터를 이용하여 Bagging, Random Forest, Xgboost, Adaboost, Light GBM으로 시행한 5겹 교차검증 결과이다. 여자 배구 경기에서는 Random Forest 의 교차검증 오분류율이 0.3672로 가장 낮았으므로 Model 2의 여자 배구 데이터에 대해서는 Random Forest를 최종 모형으로 선택하였으며, 최적 파라미터는 mtry = 2이다. Table 11는 Model 2의 여자 배구 승부 예측 최종 모형으로 선택한 Decision Tree를 사용하여 시험데이터의 승패를 예측한 결과이며, 예측 정확도는 0.6497이다. 혼동행렬을 보았을 때 실제 경기 결과와 다르게 분류한 경우가 승리와 패배 여부에 따라 큰 차이를 보이지 않는 것으로 나타났다.

3.3. 한국 프로 배구 최종 승패 예측 모형

데이터 구조를 달리하여 남자 배구와 여자 배구의 경기 승부 예측 모형을 설계한 결과, 남자 배구 경기의 예측 성능은 Model 1에서 0.572, Model 2에서 0.5396이었으며 여자 배구 경기의 예측 성능은 Model 1에서 0.5791, Model 2에서 0.6497인 것으로 나타났다. 따라서 남자 배구 최종 승패 예측 모형으로는 각 팀을 기준으로 정렬한 데이터(Model 1)를 랜덤포레스트로 예측한 모델을 선택하였으며, 여자 배구 최종 승패 예측 모형으로는 각 경기 결과를 홈팀을 기준으로 변형한 데이터(Model 2)를 랜덤포레스트로 예측한 모델을 선택하였다. 또한 본 연구에서는 최종으로 선택한 모형에 대한 변수 중요도 그림과 부분 의존도 그림을 통해 각 변수가 승리 확률과 가지는 관계와 기여도를 파악하였다.

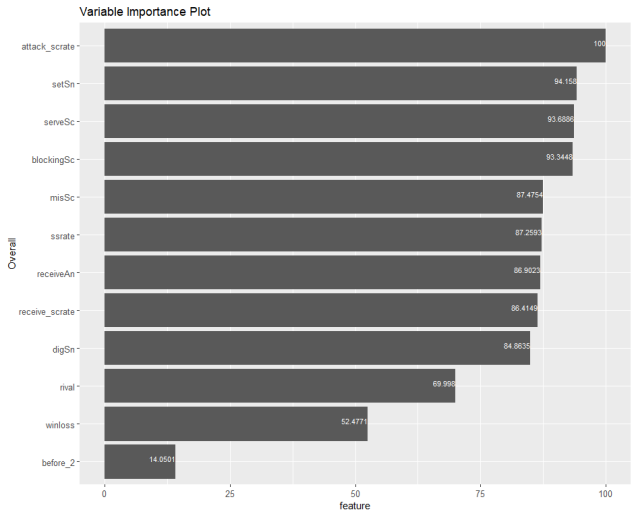

Figure 4: *Importance of predictive variables in Men's volleyball matches.*

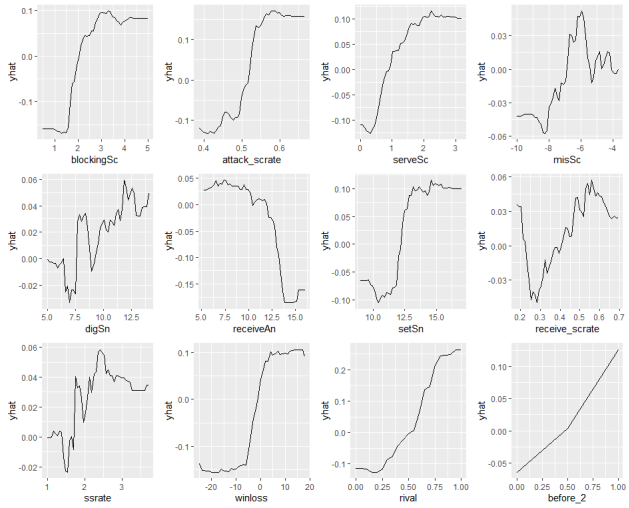

Figure 5: *Partial dependence plots for Men's volleyball matches.*

3.3.1. 남자 프로 배구 리그의 최종 승패 예측 모형 (Model 1)

Figure 4는 남자 배구 최종 승패 예측 모형에 대한 변수 중요도 그림이다. 공격성공률, 세트 성공 수, 서브 득점, 블로킹 득점, 범실 개수가 상위 5개 중요한 변수이며, 상대팀과의 전적과 관련된 변수인 상대 전적, 연승연패, 직전 2경기 승률은 상대적으로 중요도가 낮은 것으로 나타났다. 전체 득점 중 큰 비중을 차지하지 않는 블로킹 득점과 서브 득점의 중요도가 높은 점을 미루어 보았을 때 남자 배구 경기에서는 공격뿐만 아니라 공격 외 득점이 특히 승리에 기여하는 역할이 크다고 해석할 수 있다. 한편 본 연구에서 생성한 파생 변수 중 상대 전적, 연승연패, 직전 2경기 승률의 중요도가 높지 않으므로 남자 배구 경기에서는 이전 경기 내용이 당일 경기에 크게 영향을 미치지 않는다는 것을 알 수 있다.

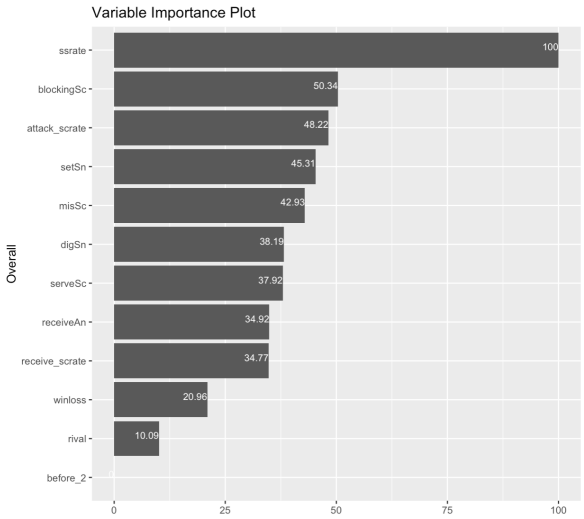

Figure 6: *Importance of predictive variables in Women's volleyball matches.*

Figure 5는 남자 배구 최종 승패 예측 모형에 대한 부분 의존도 그림이다. 대부분의 변수들은 값이 커질수록 승리 확률이 상승하는 정비례 관계를 가지는 모습을 확인할 수 있다. 한편, 리시브와 관련된 변수인 리시브 정확과 리시브 성공률이 승리 확률과 상반된 양상을 보이는 것으로 나타났다. 이를 통해 리시브를 정확히 올리는 것도 중요하지만 리시브 실패 개수를 줄이는 것이 승리 확률을 높일 수 있다고 추론할 수 있다.

3.3.2. 여자 프로 배구 리그의 최종 승패 예측 모형 (Model 2)

Figure 6은 여자 배구 최종 승패 예측 모형에 대한 변수 중요도 그림이다. 상위 5개 변수는 점유율 대비 공격성공률, 블로킹 득점 차이, 공격 성공률 차이, 세트 성공 수 차이, 범실 개수 차이다. 상대팀과의 전적과 관련된 변수인 상대 전적, 연승연패, 직전 2경기 승률은 남자 배구와 마찬가지로 상대적으로 중요도가 낮은 것으로 나타났다. 남자 배구 경기와는 달리 세트 성공수 차이, 디그 성공 수 차이가 상위권에 위치한 것으로 보아 여자 배구에서는 득점 자체보다는 득점으로 이어질 수 있는 좋은 세트와 수비가 중요하다는 것을 알 수 있다.
Figure 7는 여자 배구 최종 승패 예측 모형에 대한 부분 의존도 그림이다. 남자 배구와 마찬가지로 대부분의 변수들은 값이 커질수록 홈팀의 승리 확률이 상승하는 정비례 관계를 가지는 모습을 확인할 수 있다. 또한 리시브와 관련된 변수인 리시브 정확과 리시브 성공률이 승리 확률과 상반된 양상을 보이는 것으로 보아 여자 배구 경기에서도 리시브 실패 개수를 줄이는 것이 승리 확률에 더 크게 기여한다는 것을 알 수 있다.

4. 결론

본 연구는 2012/2013 시즌부터 2022/2023 시즌까지의 한국 배구 프로 리그 경기에 대해 분석하였다. 과거 연구들을 살펴보면, 스포츠 종목의 경기력 관련 분석들은 선수 또는 팀의 평가, 경기의 환경적 요인 평가, 전략 평가, 득점 모형, 경기 방식의 구조적 현상을 평가하기 위한 목적으로 다양한 통계적 모형이 사용되어 왔다 (Eom 등, 2002). 이처럼 통계적 모형으로는 분석의 목적이나 변수 간의 관계, 자료의 특성에 따라 다양한 방법이 사용될 수 있는데, 본 연구에서는 스포츠 경기를 구성하는 여러 가지 변인들을 고려하여 경기의 승패를 예측하는데 머신러닝 기법이 가장 적합하다고 판단하여 머신러닝을 활용한 배구 경기 승부 예측 모형을 설계하였다. 또한, 여자 배구 리그와 남자 배구 리그에 대해 각각 다른 승부 예측 모형을 설계하여 성별에 따라

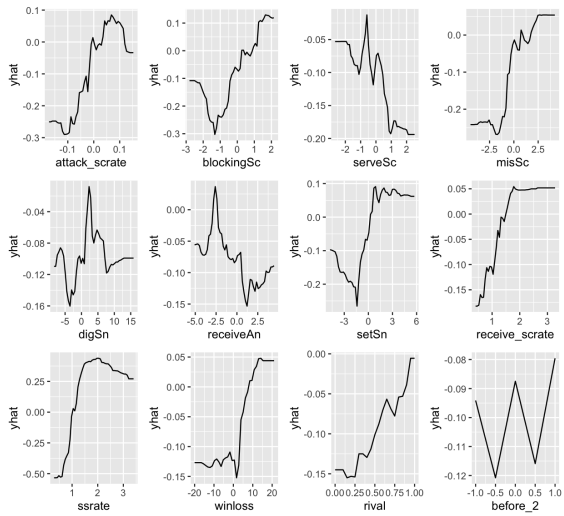

Figure 7: *Partial dependence plots for Women's volleyball matches.*

어떤 요인이 경기 승리에 더 큰 영향을 미치는지 비교하였으며, 데이터 구조를 달리하여 분석을 다각화하였다. 분석에 사용한 데이터는 KOVO 공식 사이트에서 크롤링 하였으며, Python과 R을 활용하여 데이터 전처리 및 모델링을 수행하였다. 수집한 데이터는 각 팀의 경기 결과와 각 경기에서 홈팀 기준 경기 결과를 기준으로 구조화하여 두 가지 모델로 분석을 진행하였다. 특히 경기 시작 전에는 당일 경기 내용이 결측값과 같으므로 과거 시점의 경기 내용들을 가중평균하여 업데이트한 값을 변수로 사용하였으며, 사용한 모델은 Bagging (Brieman 등, 1984), Random Forest (Breiman, 2001), Xgboost (Chen과 Guestrin, 2016), Adaboost (Friedman 등, 2000), Light GBM이다. 예측 모형 설계 결과, 남자 배구 프로 리그에서는 각 팀을 기준으로 정리한 데이터 (Model 1)를 활용한 랜덤 포레스트가 정분류율 0.572로, 여자 배구 프로 리그에서는 각 경기 결과를 홈팀을 기준으로 정리한 데이터 구조(Model 2)를 활용한 랜덤포레스트가 정분류율 0.6497로 각각 가장 높은 예측 성능을 보이는 모형인 것으로 나타났다. 최종으로 선택한 각각의 모형 대해 변수 중요도와 부분 의존도 그림을 출력하여 변수의 영향을 살펴본 결과, 성별에 따라 배구 경기 승리 확률에 기여도가 큰 변수가 다른 것을 확인할 수 있었다.

본 연구의 의의 및 기대효과는 다음과 같다. 먼저, 배구 종목을 대상으로 한 기존의 연구들이 한 가지의 성별에 한정되어 있었다면 본 연구는 여자 리그와 남자 리그를 모두 분석하여 성별에 따라 승부에 영향을 미치는 요인에 차이가 있는지 확인하였으며, 스포츠 경기에서 전통적으로 사용되는 데이터 구조를 변형하여 다양한 분석을 시도하였다. 또한 경기 내용 관련 변수 외에도 직전 경기 이력, 상대 팀과의 전적과 같은 변수를 추가로 생성하여 경기의 특성을 세분화하였다. 이를 통해 경기 특성마다 어떠한 요인이 경기를 승리로 이끄는지 파악하고 부족한 점을 보완하는 등 향후 구단 감독과 코치진들이 전략을 수립하는데 도움이 될 수 있을 것이다. 더불어, 과거 경기 내용 기록만을 활용하여 경기의 승부를 예측했음에도 불구하고 정분류율이 약 0.6에 달하는 것을 확인하였다. 마지막으로 기존 승부 예측 모형과는 다르게 해당 경기가 끝 난 이후 나오는 결과를 이용한 것이 아니라 과거의 기록들을 이용해 해당 경기의 결과를 예측한 모형이라는 점에서, 사후 예측이 아닌 사전 예측이 가능한 모형이라는 점에서 차별점을 시사한다. 이를 통해 본 연구가 스포츠 경기 승부 예측 관련 연구 발전의 계기가 될 것이라 기대한다.

References

Baacke H(1982). Statistical match analysis for evaluation of players and teams performances, *Volleyball Technical Journal*, **7**, 45–56.

Breiman L (1996). Bagging predictors, *Machine Learning*, **24**, 123–140.

Breiman L, Friedman J, Olshen R, and Stone C (1984). *Classification and Regression Trees*, Chapman and Hall, New York.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Chen T and Guestrin C (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, 785–794.

Cho SR (2017). Analysis of volleyball serve technique success rate (Master's thesis), Mokpo University, Mokpo.

Chun Y-J and Kim K-T (2011). Analysis on contents and changes of female professional volleyball's score in rally point scoring system game, *Korean Journal of Sports Science*, **20**, 729–737.

Chun Y-J and Kim K-T (2011). Analysis on contents and changes of female professional volleyball's score in rally point scoring system game, *Korean Journal of Sports Science*, **20**, 729–737.

Eom HJ and Schutz RW (1992). Statistical analyses of volleyball team performance, *Research Quarterly for Exercise and Sport*, **63**, 11–18.

Eom HJ, Jo JH, and Sin SY (2002). Application Cases of Statistical Models in Professional Sports Settings. In *Proceedings of the Korean Statistical Society Conference*, 51–59.

Eom HJ and Schutz RW (1992). Transition play in team performance of volleyball : A log-linear analysis, *Research Quarterly for Exercise and Sport*, **63**, 261–269.

Friedman J, Hastie T, and Tibshirani R (2000). "Additive logistic regression: A statistical view of boosting.", *The Annals of Statistics*, **28**, 337–407.

Hastie TJ and Pregibon D (1992). Generalized linear models. In Chambers JM and Hastie TJ (Eds), *Statistical Models in S*, Wadsworth and Brooks/Cole, Pacific Grove.

Heo C-K and Yoon J-D (2023). Bayesian Bradley-Terry with MCMC for the prediction of volleyball results, *Korean Journal of Sports Science*, **32**, 813–823.

Hong SJ, Lee KC, Kim WK, and Jang JH (2011). The development of record factor norm for evaluating tennis players, *The Korean Journal of Measurement and Evaluation in Physical Education and Sports Science*, **13**, 89–101.

Hughes M and Franks IM (1997). *Notational Analysis of Sport*, E & FN SPON, London.

Ji MJ (2014). A study on the match fixing case of domestic professional sports, *Journal of Korea Entertainment Industry Association*, **31**, 109–116.

Jo HM (1999). The impact of setter's position-based toss type and attack success rate on volleyball match outcomes (Master's thesis), College of Education, Kyung Sung University, Seoul.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, and Liu TY (2017). Lightgbm: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, **30**, 3146–3154.

Kim KW (2007). A comparative analysis of spike success rates by court and zone for high school boys' volleyball teams (Master's thesis), Mokpo University, College of Education, Mokpo.

Kim J, Kim JH, Kim J, and Hong SJ (2011). An analysis of different attacks regarding to serve receive between teams in volleyball, *Journal of Sport and Science*, **22**, 2119–2131.

Kim S (2009). A comparative analysis of success, failure, and scoring rates in male volleyball matches by serve type (Master's thesis), Busan National University of Education, Busan.

Kwon T-W, Cho S-W, and Cho Y-H (1998). The technical analysis on 's volleyball game -with the focus of women's team-, *Korean Journal of Sports Science*, **7**, 425–431.

McCullagh P and Nelder JA (1989). *Generalized Linear Models, 37*, CRC press, Boca Raton, Florida.

Shin SH (2017). The analysis of the attack type on the professional women's volleyball (Doctoral dissertation), Hanyang University, Seoul.

Received September 3, 2023; Revised October 19, 2023; Accepted November 1, 2023

한국 프로배구 연맹의 경기 예측 및 영향요인 분석

김희숙[a], 이나경[a], 이지윤[a], 송종우[1,a]

[a]이화여자대학교 통계학과

요 약

본 연구는 한국 프로배구 리그를 체계적으로 분석하고 대표적인 머신러닝 분류 방법을 활용하여 경기 결과를 예측하고자 한다. 이를 위해 2012/2013 시즌부터 2022/2023 시즌까지의 남자 프로배구와 여자 프로배구 리그 경기 데이터를 수집하였으며, 이 데이터는 경기 세부 내용을 상세하게 포함하고 있다. 데이터는 각 경기를 두 팀으로 분리한 경우와 홈팀을 기준으로 상대팀과의 성과 차이로 데이터를 가공한 경우로 두 가지 다른 데이터 구조를 모델에 적용했다. 이를 통해 남자 프로배구와 여자 프로배구 각각에 대해 총 4개의 예측 모형을 구축했다. 경기 종료 전에는 모형에서 사용하는 세부 변수 값들을 알 수 없기 때문에, 오늘 경기 직전까지의 3~4 경기의 결과를 전처리하여 이를 변수로 사용했다. 본 연구에서는 Decision Tree, Logistic Regression, Bagging, Random Forest, Xgboost, Adaboost, Light GBM 같은 다양한 머신러닝 기법을 분류에 활용하여, Random Forest를 사용한 모델이 가장 우수한 예측 성능을 보였다. 최종 선택한 모형에 대해 변수 중요도 그림과 부분 의존도 그림을 확인한 결과 성별과 데이터 구조에 따라 중요한 변수들이 다른 것으로 나타났지만, 공통적으로 세트 성공 수, 블로킹 득점, 범실 개수가 가장 중요한 변수임을 알 수 있었다. 본 승패 예측 모델은 사후적 예측이 아닌 경기 종료 전 사전 예측이 가능한 모형이라는 점에서 차별성을 가지며, 우리의 분석이 한국 프로배구 팀들에게 전략적 추론이 될 수 있을 것이라 기대한다.

주요용어: 배구, 승부예측, 머신러닝, 분류

[1]교신저자: (03760) 서울특별시 서대문구 이화여대길 52,이화여자대학교 통계학과. E-mail: josong@ewha.ac.kr