



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

석사학위논문

분류 머신러닝을 활용한 잉글랜드 프리미어리그 승패예측모형 탐색

Exploring England Premier League Win / Loss Prediction Model
using Classification Machine Learning

한국체육대학교 대학원

체육학과

이 승 박

지도교수 박 재 현

2021년 2월

분류 머신러닝을 활용한 잉글랜드 프리미어리그 승패예측모형 탐색

Exploring England Premier League Win / Loss Prediction Model
using Classification Machine Learning

한국체육대학교 대학원

체육학과

이 승 박

이 논문을 석사학위 논문으로 제출함

지도교수 박 재 현

2021년 2월

이 승 박의 석사 학위 논문을 인준함

심사위원장

윤석훈 (인)

심사위원

윤희준 (인)

심사위원

박재현 (인)

한국체육대학교 대학원

2021년 2월

국문요약

분류 머신러닝을 활용한 잉글랜드 프리미어리그 승패예측모델 탐색

한국체육대학교 대학원

체육학과

이 승 박

이 연구의 목적은 잉글랜드 프리미어리그 경기기록과 배당률 자료를 활용하여 분류 머신러닝 알고리즘 기반 축구 승패예측모형을 탐색 및 비교하는 것이다. 이 연구의 목적을 달성하기 위한 연구내용은 크게 세 부분으로 나누어진다. 첫 번째 연구내용은 ‘경기기록을 활용한 분류 머신러닝 기반 승패예측모형 탐색 및 비교’이다. 이를 위해 잉글랜드 프리미어리그 09-10시즌부터 18-19시즌 3,800경기에 대한 경기기록 자료를 잉글랜드 프리미어리그 공식 사이트(www.premierleague.com)와 후스코어드닷컴(whoscored.com)에서 수집하였으며, 분석에 사용된 변인 44개를 선정하였다. 입력 변인은 예측하려는 경기의 이전 경기 평균값을 사용하였으며, 효과적인 조합을 알아보기 위하여 이전 1~5경기에 대해서 모두 실험하고 비교하였다. 두 번째 연구내용은 ‘배당률을 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교’이다. 이를 위해 [football-data](http://football-data.com)에서 경기 배당률을 수집하였으며, 분석에 사용된 변인의 개수는 36개로 선정하였다. 세 번째 연구내용은 ‘혼합자료를 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교’이다. 이를 위해 앞서

수집한 경기기록 자료와 배당률 자료를 모두 통합하였으며, 공통변인을 포함한 변인 83개를 활용하여 승패예측모형을 탐색하였다. 이 연구의 결론은 다음과 같다.

첫째, 원자료 경기기록과 차원축소 경기기록을 활용한 승패예측모형을 탐색한 결과 차원축소 경기기록 자료와 랜덤 포레스트 알고리즘을 함께 사용한 승패예측모형(RM2)이 분류정확도(Accuracy) 54.8%로 가장 높은 순위로 나타났다.

둘째, 원자료 배당률과 차원축소 배당률을 활용한 승패예측모형을 탐색한 결과 차원축소 배당률 자료와 랜덤 포레스트 알고리즘을 함께 사용한 승패예측모형(RM2)이 분류정확도 56.6%로 가장 높은 순위로 나타났다.

셋째, 원자료 혼합자료와 차원축소 혼합자료를 활용한 승패예측모형을 탐색한 결과 혼합자료와 랜덤 포레스트 알고리즘을 함께 사용한 승패예측모형(M2)이 분류정확도 57.8%로 가장 높은 순위로 나타났다.

결론적으로 이 연구에서는 잉글랜드 프리미어리그 경기기록과 배당률 자료를 활용하여 분류 머신러닝 알고리즘 기반 축구 승패예측모형을 탐색 및 비교하였다. 이는 축구 승패예측모형을 구축할 때 자료 형태에 따른 적절한 분류 머신러닝 알고리즘 선택의 기초정보로 활용될 수 있을 것으로 기대한다.

주요어: 머신러닝, 잉글랜드 프리미어리그, 승패예측모형, 분류 알고리즘

목 차

I. 서론	1
1. 연구의 필요성	1
2. 연구내용	5
II. 이론적 배경	7
1. 분류(Classification)의 개념	7
2. 예측모형 탐색	7
3. 분류 머신러닝 알고리즘	10
III. 연구방법	17
1. 연구절차	17
2. 연구자료 및 수집방법	18
3. 자료처리방법	21
1) 전처리과정	21
(1) 과거 경기에 따른 사전변인	22
(2) 결측값 처리	22
(3) 표준화	23
(4) 범주형 변인 변환	23
2) 교차검증(Cross Validation)	24
3) 모형평가	25
4) 차원축소(Dimensionality Reduction)	27

IV. 연구결과	29
1. 경기기록을 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교	29
1) 경기기록을 활용한 승패예측모형 탐색	29
2) 차원축소 경기기록을 활용한 승패예측모형 탐색	35
3) 경기기록 승패예측모형과 차원축소 경기기록 승패예측모형 비교	40
2. 배당률을 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교	41
1) 배당률을 활용한 승패예측모형 탐색	41
2) 차원축소 배당률을 활용한 승패예측모형 탐색	46
3) 배당률 승패예측모형과 차원축소 배당률 승패예측모형 비교	51
3. 혼합자료를 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교	52
1) 혼합자료를 활용한 승패예측모형 탐색	52
2) 차원축소 혼합자료를 활용한 승패예측모형 탐색	57
3) 혼합자료 승패예측모형과 차원축소 혼합자료 승패예측모형 비교	62
V. 논의	63
VI. 결론	67
참고문헌	69
ABSTRACT	76
부록	79

표 목차

표 1. 경기기록 변인	19
표 2. 배당률 변인	20
표 3. 자료처리 및 분석에 사용된 Python 라이브러리	21
표 4. 결측값 처리 변인	23
표 5. 경기기록에 따른 의사결정나무 승패예측모형 타당도	30
표 6. 경기기록에 따른 랜덤 포레스트 승패예측모형 타당도	31
표 7. 경기기록에 따른 XGBoost 승패예측모형 타당도	31
표 8. 경기기록에 따른 LightGBM 승패예측모형 타당도	32
표 9. 경기기록에 따른 로지스틱 회귀분석 승패예측모형 타당도	33
표 10. 경기기록에 따른 서포트 벡터 머신 승패예측모형 타당도	33
표 11. 경기기록에 따른 최적 승패예측모형 타당도 비교	34
표 12. 차원축소 경기기록에 따른 의사결정나무 승패예측모형 타당도	35
표 13. 차원축소 경기기록에 따른 랜덤 포레스트 승패예측모형 타당도	36
표 14. 차원축소 경기기록에 따른 XGBoost 승패예측모형 타당도	37
표 15. 차원축소 경기기록에 따른 LightGMB 승패예측모형 타당도	37
표 16. 차원축소 경기기록에 따른 로지스틱 회귀분석 승패예측모형 타당도	38
표 17. 차원축소 경기기록에 따른 서포트 벡터 머신 승패예측모형 타당도	39
표 18. 차원축소 경기기록에 따른 최적 승패예측모형 타당도 비교	39
표 19. 최적 승패예측모형 종합 타당도 및 순위 비교	40
표 20. 배당률에 따른 의사결정나무 승패예측모형 타당도	41
표 21. 배당률에 따른 랜덤 포레스트 승패예측모형 타당도	42
표 22. 배당률에 따른 XGBoost 승패예측모형 타당도	43
표 23. 배당률에 따른 LightGBM 승패예측모형 타당도	43
표 24. 배당률에 따른 로지스틱 회귀분석 승패예측모형 타당도	44

표 25. 배당률에 따른 서포트 벡터 머신 승패예측모형 타당도	45
표 26. 배당률에 따른 최적 승패예측모형 타당도 비교	45
표 27. 차원축소 배당률에 따른 의사결정나무 승패예측모형 타당도	46
표 28. 차원축소 배당률에 따른 랜덤 포레스트 승패예측모형 타당도	47
표 29. 차원축소 배당률에 따른 XGBoost 승패예측모형 타당도	48
표 30. 차원축소 배당률에 따른 LightGBM 승패예측모형 타당도	48
표 31. 차원축소 배당률에 따른 로지스틱 회귀분석 승패예측모형 타당도	49
표 32. 차원축소 배당률에 따른 서포트 벡터 머신 승패예측모형 타당도	50
표 33. 차원축소 배당률에 따른 최적 승패예측모형 타당도 비교	50
표 34. 최적 승패예측모형 종합 타당도 및 순위 비교	51
표 35. 혼합자료에 따른 의사결정나무 승패예측모형 타당도	52
표 36. 혼합자료에 따른 랜덤 포레스트 승패예측모형 타당도	53
표 37. 혼합자료에 따른 XGBoost 승패예측모형 타당도	54
표 38. 혼합자료에 따른 LightGBM 승패예측모형 타당도	54
표 39. 혼합자료에 따른 로지스틱 회귀분석 승패예측모형 타당도	55
표 40. 혼합자료에 따른 서포트 벡터 머신 승패예측모형 타당도	56
표 41. 혼합자료에 따른 최적 승패예측모형 타당도 비교	56
표 42. 차원축소 혼합자료에 따른 의사결정나무 승패예측모형 타당도	57
표 43. 차원축소 혼합자료에 따른 랜덤 포레스트 승패예측모형 타당도	58
표 44. 차원축소 혼합자료에 따른 XGBoost 승패예측모형 타당도	59
표 45. 차원축소 혼합자료에 따른 LightGBM 승패예측모형 타당도	59
표 46. 차원축소 혼합자료에 따른 로지스틱 회귀분석 승패예측모형 타당도	60
표 47. 차원축소 혼합자료에 따른 서포트 벡터 머신 승패예측모형 타당도	61
표 48. 차원축소 혼합자료에 따른 최적 승패예측모형 타당도 비교	61
표 49. 최적 승패예측모형 종합 타당도 및 순위 비교	62

그림 목차

그림 1. CRISP-DM 도식화	8
그림 2. 배깅(bagging) 절차	11
그림 3. 부스팅(boosting) 절차	12
그림 4. 서포트 벡터 머신	15
그림 5. 연구절차	17
그림 6. 사전변인 전처리과정	22
그림 7. OneHotEncoder 방법	24
그림 8. 어웨이팀 삼원 분류표 예시	26
그림 9. 경기결과 예측 삼원분류표	64

수식

<수식 1>	13
<수식 2>	14
<수식 3>	14
<수식 4>	14
<수식 5>	14
<수식 6>	26
<수식 7>	27
<수식 8>	27
<수식 9>	27

I. 서론

1. 연구의 필요성

인간은 생존과 발전을 위해 현재를 이해하고 미래를 예측하는 노력을 지속하고 있다. 그 결과 미래를 탐구하는 학문이 생겨났으며, 이러한 학문은 ‘과학기술을 기반으로 한 예측’에서 비롯되었다(Linstone, 2002). 과학기술을 바탕으로 미래를 예측하려는 시도는 다양한 분야에서 이루어지고 있다. Chui(2017)에 따르면 헬스케어 분야에서 사전에 질병을 예측함으로써 환자의 위험을 낮추고 미국 국민의 의료비용을 5~9% 감소시켰다. 제조업에서는 물품들에 대한 수요를 예측함으로써 배송 효율을 30% 이상 증가시켰으며, 사용 연료를 약 12% 감소시켰다.

미래를 예측하려는 시도는 스포츠를 연구하는 분야에서도 찾아볼 수 있다. 구체적으로 시계열 분석을 통한 프로스포츠 관중 수 예측(김혁, 2019; 최재일, 정용락, 2010), 선수 연봉 예측모형 개발 및 예측 변인 탐색에 관한 연구(김세형, 강상조, 박재현, 김혜진, 2008; 이장택, 2020; 한필수, 이석인, 2003)등이 이루어지고 있다. 이처럼 스포츠 현장에서 미래를 예측하려는 시도는 다양한 관점에서 진행 중이며, 그중 많은 변인이 상호작용하는 경기결과 예측은 스포츠 현장뿐만 아니라 중요한 연구문제로 주목받고 있다(Arabzad, Tayebi Araghi, Sadi-Nezhad, & Ghofrani, 2014; Owranipur, Eskandarian, & Mozneb, 2013).

스포츠 현장에서 경기결과를 예측하는 것은 감독, 선수와 같은 직접적인 관계자들뿐만 아니라 언론, 팬, 스포츠 기업 역시 많은 관심을 보인다. 매년 언론에서는 챔피언스리그 우승팀, 월드컵 우승 선수, 내셔널 풋볼 리그 우승팀 등을 예상하는 프로그램들을 방영하고 있으며, 열성적인 팬들은 각 구단의 전술, 선수들의 장단점 등을 종합적으로 평가 및 분석하여 의견을 제시하고 블로그에 누가 승리할지에 대한 토론의 장을 마련하고 있다. 경기결과를 예측하는 스포츠 팬의 증가 현상은 경기결과 예측 성공 여부에 따라 배당금을 지급하는 배팅 사업의 확장요인으로 작용하였으며(채진석, 조은형, 엄한주, 2010), 배팅 회사들도

적절한 배당률을 책정하기 위해서 경기결과를 예측하고 분석하고 있다(Spann & Skiera, 2009). 경기결과를 예측하는 연구는 다양한 학문 분야에서 진행되어오고 있다. 스포츠 분야 외에 공학, 정보통신, 데이터 과학 분야에서도 경기결과 변인들을 이용하여 프로스포츠 승패를 예측하는 연구(이석원, 천영진, 2017; 이재현, 이수원, 2020; Joseph, Fenton, & Neil, 2006; Min, Kim, Choe, Eom, & Mckay, 2008)가 시도되어오고 있다.

축구는 경기가 진행되는 동안 다양한 위치에서 끊임없이 이벤트가 발생하는 스포츠이다. 이러한 특성 때문에 스포츠과학자들은 축구를 과학적이고 체계적으로 기록과 분석을 하고 있으며(최형준, 이운수, 2019), 축적된 자료들은 경기의 상황을 이해하는 것에서 그치지 않고 경기결과를 예측하는 자료로 활용된다. 특히 매년 평균 3만 명 이상의 관중 규모를 유지하는 잉글랜드 프리미어리그는 일찍이 축구에서 발생하는 자료를 활용하여 결과를 예측함으로써 사람들에게 다양한 정보를 제공하고 있다(Domingues, Lopes, Mihaylova, & Georgieva, 2019).

분석적인 측면에서 축구 경기결과를 예측하는데 적용되고 있는 방법은 다양하게 시도되어왔으며, 대표적으로 포아송 분포를 적용한 것을 확인할 수 있었다. 포아송 분포를 활용한 경기결과 예측모형을 살펴보면 축구경기에서 각 팀은 득점을 이용한 공격력과 방어력에 따른 각각의 포아송 분포를 가진다는 전제로 팀의 승률을 예측하고 있다. 포아송 분포를 활용해서 경기결과를 예측하는 연구는 1982년을 시작으로 계속해서 연구되어오고 있다(Dixon & Coles, 1997; Maher, 1982; Rue & Salvesen, 2000).

2016년 알파고의 등장 이후 자주 사용되고 있는 분석방법은 머신러닝(Machine Learning)이다(박홍진, 2020). 머신러닝은 인공지능(AI: artificial intelligence)의 한 분야로써 컴퓨터가 자료를 통해 직접 예측모형을 구현하는 기술이다(Tiwari, Tiwari, & Tiwari, 2018). 머신러닝을 활용한 경기결과 예측에는 주로 분류 머신러닝 알고리즘이 사용된다. 분류 머신러닝은 지도학습의 한 종류로서 주어진 자료를 정해진 카테고리에 따라 분류하는 알고리즘이다(Kotsiantis, Zaharakis, & Pintelas, 2007). 축구에서는 승과 패로 분류하거나, 승, 무, 패로 분류하는 역할을 하며, 축구 경기결과 예측에 사용되는 대표 알고리즘은 로지스틱 회귀분

석, 서포트 벡터 머신, 의사결정나무, 랜덤포레스트, XGBoost, LightGBM 등이 있다 (Hubáček, Šourek, & Železný, 2019; Iskandaryan, Ramos, Palinggi, & Trilles, 2020; Prasetyo, 2016).

분류 알고리즘을 활용한 축구 승패예측모형과 관련된 연구는 국내외적으로 계속 진행되고 있다. 김형원(2020)은 잉글랜드 프리미어리그 12개의 시즌을 대상으로 인공신경망과 XGBoost를 사용하여 모형을 구축하였다. 인공신경망은 53.9%, XGBoost는 58.2%의 정확도를 보였다. 해외에서는 분류 알고리즘을 이용한 승패예측모형을 구축한 연구(Buursma, 2011; Igiri & Nwachukwu, 2014; Nivard & Mei, 2012)와 선행연구의 축구 승패예측모형을 비교하는 연구(Gevaria, Sanghavi, Vadiya, & Deulkar, 2015)가 진행되어왔다. 하지만 분류 알고리즘을 사용한 축구 승패예측모형 연구를 살펴보면 계산 복잡성, 저장공간과 같은 모형의 효율성을 높이기 위한 노력을 찾아보기 쉽지 않았다. 또한, 승패예측모형을 비교한 연구에서는 각 연구에서 사용한 학습자료가 다름에도 불구하고 똑같은 기준으로 비교하였으며, 모형의 타당도를 정밀도, 재현도, F-1 score과 같이 다양한 지수를 고려하지 않고 단순히 정확도만을 사용하였다는 점에서 최적 모형을 정확하게 판단할 수 없었다.

종합적으로 분류 머신러닝 알고리즘을 활용한 축구 승패예측모형과 관련된 선행연구에서 몇 가지 한계점을 확인하였다. 첫째, 자료의 특성에 맞는 효과적인 알고리즘을 결정하는데 많은 시행착오를 겪는다는 문제점이 있다(이영섭, 오현정, 김미경, 2005). 축구 승패예측모형과 관련된 선행연구들의 자료 형태는 경기에서 발생한 이벤트를 기록한 경기기록자료(Stübinger, Mangold, & Knoll, 2020), 배팅 회사들의 배당률(Štrumbelj & Šikonja, 2010; Wunderlich & Memmert, 2018), 마지막으로 두 가지 자료를 함께 사용한 혼합자료(Hoekstra, Bison, & Eiben, 2012; Tax & Joutstra, 2015)로 나누어진다. 하지만 자료의 형태에 따라서 연구자들은 어떤 분류 알고리즘을 사용해야 할지 판단을 내리기 어려우며, 더욱이 효율적인 분류 알고리즘을 선택하지 못할 경우 많은 시간과 노력을 낭비하고 얻어진 결과 또한 신뢰성을 잃는다.

둘째, 분류 머신러닝 알고리즘을 활용해서 예측모형을 구성할 때 나타나는 문제는 경제

성의 원리(Principle of economy)를 고려하지 않는다는 것이다. 경제성의 원리란 오컴의 면도날(Occam's Razor)이라고 불리며 어떤 현상을 설명할 때 불필요한 가정을 제거하고 가장 단순하게 설명하는 것이다(Blumer, Ehrenfeucht, Haussler, & Warmuth, 1987). 머신러닝 모형이 복잡할 경우 해석이 어렵고, 학습시간이 길어지며 모형의 일반화가 잘 이루어지지 않는 과적합 현상이 발생한다. 이를 해결하기 위해서 차원축소(Dimensionality Reduction)가 강조되고 있으며, 차원축소 자료를 이용한 승패예측모형과 원자료를 이용한 승패예측모형 간의 타당도를 확인할 필요가 있다.

따라서 이 연구의 목적은 잉글랜드 프리미어리그 10개 시즌의 경기기록과 배당률 자료를 활용하여 분류 머신러닝 알고리즘 기반 축구예측모형을 탐색하는 것이다. 나아가 모형의 경제성을 고려하기 위해서 차원축소 자료를 활용하여 승패예측모형을 탐색하고, 원자료를 이용한 승패예측모형과 비교하였다. 구체적으로 첫째, 원자료 경기기록과 차원축소 경기기록을 활용해서 승패예측모형을 탐색 및 비교하였다. 둘째, 배당률 원자료와 차원축소 배당률 자료를 활용하여 승패예측모형을 탐색 및 비교하였다. 셋째, 경기기록과 배당률 자료를 합친 원자료와 차원축소 혼합자료를 활용해서 승패예측모형을 탐색 및 비교하였다.

2. 연구내용

이 연구는 잉글랜드 프리미어리그 경기기록과 배당률 자료를 활용하여 분류 머신러닝 알고리즘 기반 축구 승패예측모형을 탐색하는 것이 목적이다. 이 연구의 내용은 크게 세 부분으로 나누어진다. 첫째, 원자료 경기기록과 차원축소 경기기록을 활용한 승패예측모형들을 탐색 및 비교한다. 둘째, 원자료 배당률과 차원축소 배당률을 활용한 승패예측모형들을 탐색 및 비교한다. 셋째, 원자료 혼합자료와 차원축소 혼합자료를 활용한 승패예측모형들을 탐색 및 비교하는 것이다. 각 연구내용에서 탐색의 범위는 6가지의 분류 머신러닝 알고리즘(의사결정나무, 랜덤 포레스트, XGBoost, LightGBM, 서포트 벡터 머신, 로지스틱 회귀분석)을 활용하여 승패예측모형을 학습시키고, 각 모형의 타당도를 확인하는 것이다. 비교는 원자료를 활용한 승패예측모형과 차원축소를 활용한 모형을 종합적으로 비교하고 순위를 매김으로써 우수 모델의 자료 조합을 확인하는 것이다.

연구내용 1. 경기기록을 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교

- 1-1. 경기기록을 활용한 승패예측모형 탐색
- 1-2. 차원축소 경기기록을 활용한 승패예측모형 탐색
- 1-3. 경기기록 승패예측모형과 차원축소 경기기록 승패예측모형 비교

연구내용 2. 배당률을 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교

- 2-1. 배당률을 활용한 승패예측모형 탐색
- 2-2. 차원축소 배당률을 활용한 승패예측모형 탐색
- 2-3. 배당률 승패예측모형과 차원축소 배당률 승패예측모형 비교

연구내용 3. 혼합자료를 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교

3-1. 혼합자료를 활용한 승패예측모형 탐색

3-2. 차원축소 혼합자료를 활용한 승패예측모형 탐색

3-3. 혼합자료 승패예측모형과 차원축소 혼합자료 승패예측모형 비교

II. 이론적 배경

1. 분류(Classification)의 개념

머신러닝의 종류는 지도 학습(Supervised Learning), 비지도 학습(Unsupervised Learning), 강화 학습(Reinforcement Learning)으로 구분된다(Kotsiantis, Zaharakis, & Pintelas, 2007). 연구에서 사용된 지도학습은 실제값과 예측값의 오차를 최소화하는 모델을 자동으로 구축하는 방법이다(Muhammad & Yan, 2015). 지도학습은 입력값(input data)과 입력값에 대한 정답인 레이블(label)을 이용하여 모델을 학습시키는 방식이다. 예를 들어 동물 사진을 분류하는 모델을 구축한다고 가정할 때 동물 사진들의 픽셀(pixel)값은 입력값을 의미하며, 각 사진이 강아지인지 고양이인지 구분하는 레이블은 입력값들에 대한 정답이다. 동물 사진들에 대한 입력값과 레이블 자료를 이용해 실제 정답과 모델이 분류한 값의 오차를 최소화하는 모델을 구축하는 것이 하나의 지도학습 학습 과정을 나타낸다.

분류(Classification)는 지도학습 종류 중 하나이다. 머신러닝과 통계학에서의 분류는 새로 관측된 자료가 어떤 범주 집합에 속하는지를 식별하는 것을 의미한다(Yom-Tov, 2003). 학습자료를 이용해 모델을 학습하면, 모델은 자료를 분류하는 결정 경계(Decision boundary)를 만든다. 이 연구에서는 과거 경기기록과 배당률 자료를 통해 축구 경기결과(홈승, 어웨이승, 무승부)를 분류하는 결정 경계를 설정하여 모델을 구축하였다.

2. 예측모형 탐색

이 연구는 과거 경기기록과 배당률 자료를 활용하여 분류 머신러닝 알고리즘 기반 축구 승패예측모형을 탐색하였다. 연구에서 의미하는 탐색의 범위는 다음과 같이 설정하였다. 예측모형 탐색은 자료수집부터 모형의 효율성을 극대화 시키는 모형개선까지의 전 과정을

의미한다. 예측모형 탐색을 위한 머신러닝 체계는 CRISP-DM(Cross-Industry Standard Process for Data Mining)(Shearer, 2000)을 참고하였다. CRISP-DM은 <그림 1>과 같다.

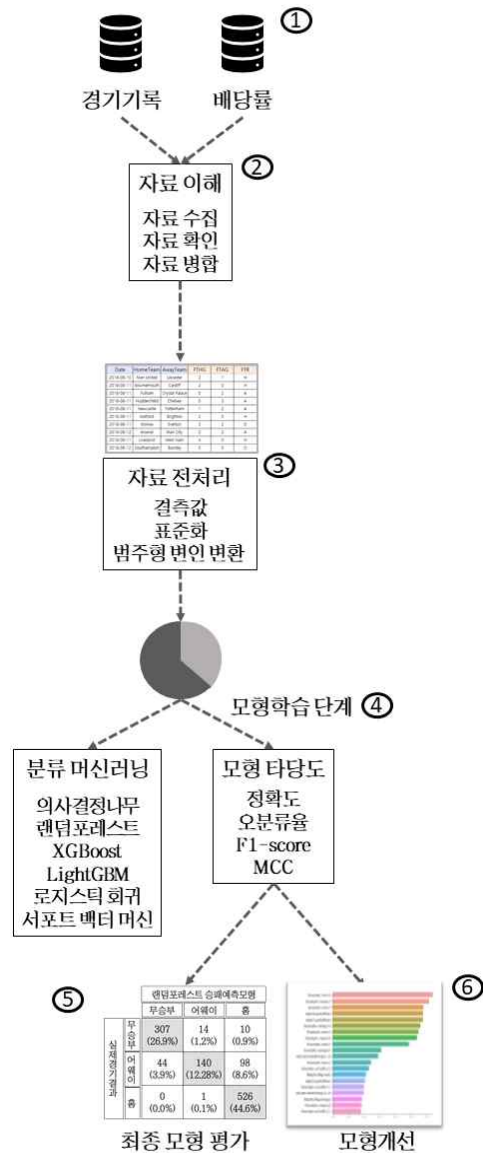


그림 1. CRISP-DM 도식화

CRISP-DM은 6단계로 나누어진다. 첫째, 관련 연구목적을 설정하고 해당 종목과 리그를 이해하는 단계이다. 둘째, 자료이해 단계이다. 이 과정은 자료를 수집하고 자료의 상태를 확인하는 단계이다. 셋째, 자료 전처리(preprocessing) 단계이다. 자료를 분석하기 위한 형태로 변환하며, 결측값, 표준화 작업등이 포함된다. 넷째, 모형학습 단계이다. 각 분류 머신러닝 알고리즘을 적용하여 모형을 학습시키고, 교차검증을 통한 모형의 타당도를 검증한다. 다섯째, 최종 모형 평가단계이다. 교차검증을 통해 최종 모형이 선택되면 검증자료로 모형을 평가한다. 마지막은 모형개선 단계이다. 실제 상황에서 사용하기 위해 모형의 문제점을 확인하고 효율적인 방향으로 개선하는 작업이 진행된다.

3. 분류 머신러닝 알고리즘

1) 의사결정나무

의사결정나무는 최적의 의사결정 규칙을 나무구조로 모형화하는 방법이다(최창환, 윤지운, 2017). 의사결정나무는 자료의 분류(classification)와 예측(prediction)을 수행하는데 유용한 방법으로 알려져 왔으며 여러 장점이 제시되어왔다(Breiman, Friedman, Stone, & Olshen, 1984). 첫째, 범주형 종속변인(승/패, 성공/실패) 분류에 효과적이며, 시각화하기 편리하다. 둘째, 의사결정나무는 비 모수적 통계기법으로 모수통계의 기본가정인 선형성, 등분산성, 분포의 정상성에 자유롭다. 셋째, 다양한 입력 변인들이 종속변인에 영향을 미치는 상호작용 효과를 파악할 수 있다.

이 연구에서는 의사결정나무의 다양한 알고리즘 중 해석이 편리하다는 장점을 가지고 있는 CART(classification and regression trees) 알고리즘을 사용하였다. CART는 부모 노드에서 자식 노드로 분리할 때 이진 분리를 하며, 정보의 불순도를 감소시키는 입력 변인을 찾고, 분리 기준점을 찾는 방식이다. 이 과정을 반복하여 하나의 나무구조를 형성한다. CART는 항상 이진 분리하기 때문에 해석이 쉽다는 장점이 있지만, 입력 변인의 개수가 적으면 예측력이 떨어진다는 것과 모형이 불안정하여 자료의 형태가 조금만 달라져도 전혀 다른 결과가 나올 수 있다는 단점이 보고되고 있다(유진은, 2015). 이러한 CART의 단점을 보완하기 위하여 정확도가 낮은 여러 의사결정나무 모형을 조합한 앙상블(Ensemble) 방법이 개발되었다. 이 연구에서는 의사결정나무의 단점을 보완한 앙상블(Ensemble) 기반 랜덤 포레스트, XGBoost, LightGBM 알고리즘을 추가로 사용하였으며, 각 알고리즘의 설명은 다음과 같다.

2) 랜덤 포레스트

의사결정나무의 단점을 보완하기 위한 앙상블 방법은 방식에 따라 배깅(bagging)과 부스팅(boosting) 방법으로 나누어진다. 배깅은 주어진 자료에 대해서 여러 개의 붓스트랩

(bootstrap) 자료를 생성하고 각 붓스트랩 자료를 분류 알고리즘을 이용하여 학습시킨 후 결합하여 최종 예측모형을 산출하는 방법이다. 구체적인 방법은 <그림 2>와 같다. 전체 자료 A에서 복원 단순 임의추출하여 n 개의 붓스트랩 자료를 생성한다. 각 붓스트랩 자료에 분류 알고리즘을 적용하여 L 개의 모형을 얻는다. L 개의 모형들을 결합하여 하나의 최종 예측모형을 만드는 방법은 종속 변인이 범주형일 경우는 다중 투표를 사용하고, 연속형일 경우 평균을 사용한다.

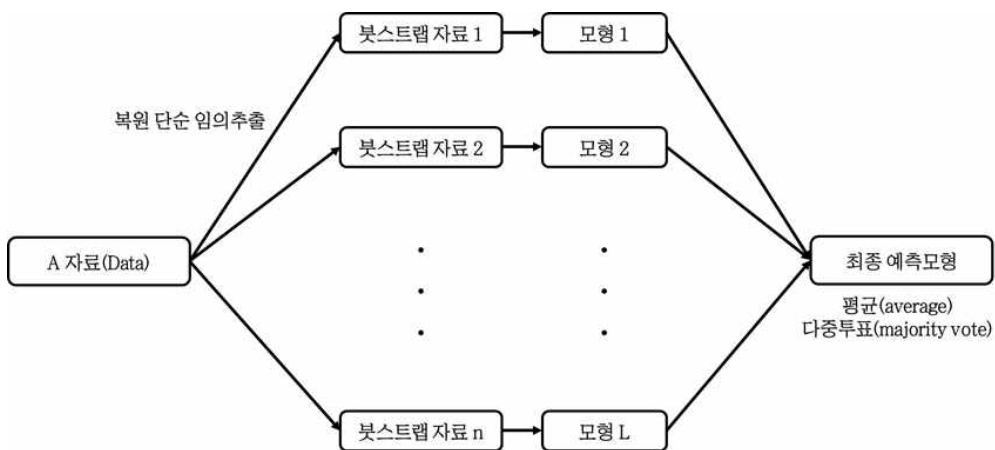


그림 2. 배깅(bagging) 절차

랜덤 포레스트는 배깅 방법을 이용한 대표 알고리즘이다. 랜덤 포레스트는 다수의 의사결정나무 모형을 만들고 종합하여 최종 예측모형을 만드는 알고리즘이다(Breiman, 2001). 의사결정나무와 배깅 방법을 결합한 랜덤 포레스트는 안정성과 정확성을 향상시켰다는 평가를 받고 있다. Breiman(2001)은 붓스트랩 자료가 무작위로 추출된 자료와 입력 변인들에 의해 만들어지므로, 대수의 법칙에 의해 모형의 예측력이 향상된다고 하였다. 또한, 많은 모형의 평균값으로 맞추기 때문에 과적합 현상을 방지할 수 있다는 장점을 강조하였다.

3) XGBoost

앙상블 방법의 다른 종류인 부스팅은 이전 모형에서 잘못 분류한 자료에 가중치를 부여하여 다음 모형에 사용함으로써 오차를 줄여나가는 방법이다. 배깅과 다른 점은 병렬처리가 불가능하다는 점이다. 배깅은 각각의 붓스트랩 자료가 독립적이기 때문에 생성된 모형들끼리 연관성이 없다. 그러므로 병렬로 자료들을 처리함으로써 속도를 증가시킨다. 반대로 부스팅은 이전 모형의 결과를 바탕으로 새로운 모형을 만드는 방법이기 때문에 학습시간이 길어지지만 정확도가 높아진다는 장점이 있다. 부스팅 방법에 대한 절차는 다음 <그림 3>과 같다.

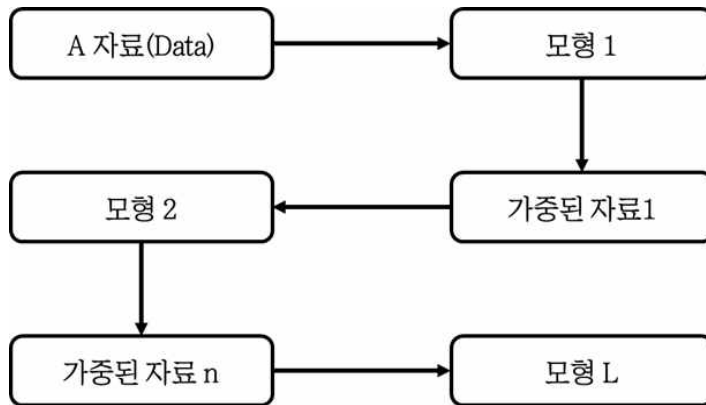


그림 3. 부스팅(boosting) 절차

분석에 사용되는 A 자료의 가중치는 동일한 상태에서 분류 알고리즘을 이용하여 모형1을 생성한다. 이때, 오분류된 자료에는 높은 가중치를 주고 올바르게 분류된 자료에는 낮은 가중치를 부여해 가중된 자료1을 생성한다. 이와 같은 과정을 반복하여 분류하기 힘든 자료를 더 잘 분류할 수 있도록 하는 최종 모형 L을 생성한다.

XGboost는 부스팅 방법을 사용한 대표 알고리즘이다(Chen, & Guestrin, 2016). XGboost는 예측값과 실제값의 차이를 나타내는 손실함수와 트리 모형의 복잡도 함수로 이루어진 목적함수를 활용하여 최적의 모형을 만든다(하대우, 김영민, 안재준, 2019). 목적함수의 공식은 <수식 1>과 같다.

<수식 1>

$$Obj = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

공식의 $l(y_i, \hat{y}_i)$ 는 실제값과 예측값의 차이를 나타내는 손실함수를 의미한다. (f_k) 는 각 모형의 나무와 잎의 복잡도 즉 모형의 복잡도를 의미하며, $\Omega(f_k)$ 는 모형의 복잡도를 조절하여 과적합 현상을 방지하는 정규화 항이다. 그리고 테일러 급수 및 1차 미분, 2차 미분을 활용하여 최적의 가중치를 구한다. 초기 부스팅 방법과 비교하여 XGBoost가 가지고 있는 장점은 첫째, 기존 부스팅 대비 학습 수행 시간이 짧다. 기존의 부스팅 방법은 한 모형의 결과에 따라서 가중치를 조절하는 방법이기 때문에 속도가 느렸지만, XGBoost는 병렬 수행과 GPU(Graphics Processing Unit) 사용이 가능해지면서 학습속도가 상승하였다. 둘째, 과적합을 스스로 규제한다. 목적함수에 나와 있는 것처럼 모형의 복잡도에 대한 규제항이 있어 초기 부스팅보다 과적합 방지에 효과적이다.

4) LightGBM

LightGBM은 XGBoost의 장점은 계승하고 단점을 보완한 부스팅 기반 알고리즘이다(Ke, et al., 2017). 기존의 부스팅 방법을 이용한 분류모형은 균형트리 분할(level wise) 방식을 이용하였지만 LightGBM은 리프중심트리분할(leaf wise) 방식을 사용한다. 리프중심트리분할 방식은 트리의 균형을 생각하지 않고 최대 손실 값을 가지는 노드를 계속 분할 한다. 이 방식은 나무의 깊이를 확장하기 때문에 모형의 정확도를 증가시킨다. 또한, 자료의 개수를 줄이는 GOSS(gradient-based one-side sampling)와 변인의 개수를 줄이는 EFB(exclusive feature bundling) 방법을 통해 메모리 사용을 감소시키고, 학습속도를 증가시켰다는 장점이 있다(김인호, 이경섭, 2020). 하지만 학습을 지속할수록 나무의 깊이가 깊어져 자료의 규모가 작은 경우 과적합 현상이 발생한다.

5) 로지스틱 회귀분석

로지스틱 회귀분석은 독립 변수의 선형 결합으로 종속 변수를 설명한다는 관점에서 선형 회귀분석과 유사하다. 하지만 선형 회귀분석과 다르게 종속 변수의 형태가 범주형 자료이다. 즉, 학습자료가 입력되었을 때 해당 자료의 결과가 특정 범주로 나누어지기 때문에 분류 알고리즘에 속한다(Bishop, 2006).

$$\langle \text{수식 2} \rangle \\ f(x) = \frac{1}{1 + e^{-x}}$$

일반적으로 회귀분석은 주어진 독립 변수에 따라 최소자승법을 이용하여 종속 변수를 예측하는 선형식이다. 하지만 로지스틱 함수는 독립 변수에 따른 사건 발생 확률로 표현된다(Hosmer Jr, Lemeshow, & Sturdivant, 2013). <수식 2>는 선형 회귀식을 확률에 관한 식으로 변형한 로지스틱 함수이다. $f(x)$ 는 0과 1사이의 값을 갖는 값이 되기 때문에 확률로 사용한다.

$$\langle \text{수식 3} \rangle \\ odds\ ratio = \frac{p}{1-p}$$

구체적인 로지스틱 함수 사용법은 다음과 같다. 로지스틱 함수의 결과값을 odds 비율로 가정한다. 여기서 odds 비율이란 p (성공확률)을 $1-p$ (실패확률)로 나눈 값을 의미한다. 위의 함수는 계수들에 대해서 비선형 함수이기 때문에 자연로그를 취하여 선형으로 변환하는데, 이 과정을 로짓 변환(logit transformation)이라고 한다. 변환된 식은 아래 <수식 4>로 표현되며, 이를 성공확률 p 에 대해 이항하면 완성된 <수식 5>를 얻을 수 있다. 성공확률에 대한 역 로지스틱 함수를 통해서 입력 자료에 따른 결과의 확률을 알 수 있다.

$$\langle \text{수식 4} \rangle \\ \ln \frac{p_i}{1-p_i} = \beta \cdot X_i$$

$$\langle \text{수식 5} \rangle \\ p_i = \frac{1}{1 + e^{-\beta \cdot X_i}}$$

6) 서포트 벡터 머신

서포트 벡터 머신은 자료를 분류하기 위한 기준선 즉 결정 경계(decision boundary)를 구하는 알고리즘이다(Cortes & Vapnik, 1995). 결정 경계가 정해지면 자료가 입력될 때 경계에 따라서 분류를 수행한다. 결정 경계는 변인이 2개일 경우 평면에서 선으로 나타나지만, 변인의 개수가 3개일 경우 선이 아닌 평면으로 나타난다. 변인의 개수가 3개 이상일 때는 단순히 평면이 아닌 더 높은 차원이 되기 때문에 초평면(hyperplane)이라고 부르게 된다. 그러므로 서포트 벡터 머신은 자료들의 초평면을 구하는 알고리즘이라고 표현되기도 한다.

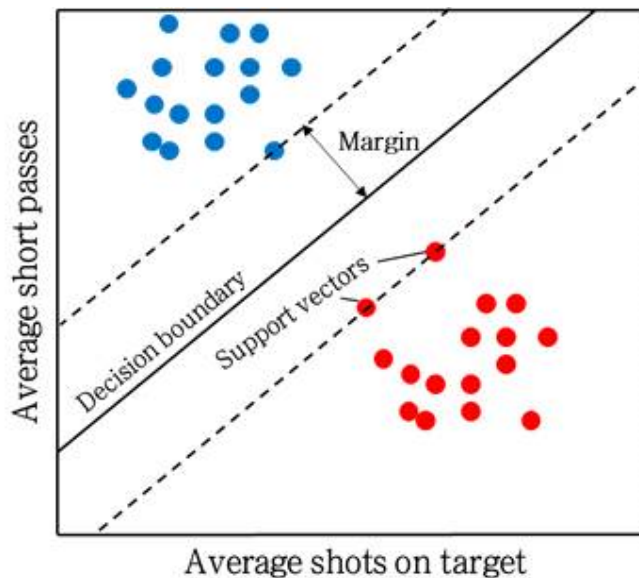


그림 4. 서포트 벡터 머신

서포트 벡터(support vectors)는 결정 경계와 가까이 있는 자료들을 의미하며, 마진(margin)은 결정 경계와 서포트 벡터 사이의 거리를 나타낸다(윤보람, 2020). 최적의 결정 경계는 자료들의 군집으로부터 최대한 멀리 떨어져 있는 것이며, 이는 마진을 최대화한 것을 의미한다. 많은 경우 선형으로 결정 경계를 결정할 수 없다. 따라서 위에서 언급한 변

인들의 개수를 늘려 높은 차원에서 분리하는 평면을 구하는 방법을 사용한다. 하지만 저차원 자료를 고차원으로 보내고 서포트 벡터를 구하고 다시 저차원으로 내리는 과정은 복잡할 뿐만 아니라 연산량이 많아진다는 단점이 있다.

Ⅲ. 연구방법

1. 연구절차

연구의 목적을 달성하기 위해서 잉글랜드 프리미어리그 공식 사이트(premierleague.com)와 후스코어드닷컴(whoscored.com)에서 제공하는 경기기록 자료를 연구자료로 수집하였으며, 배당률은 과거 배당률 자료를 제공하는 football-data에서 수집하였다. 연구절차는 7단계로 나누어지며 구체적인 절차는 아래 <그림 5>와 같다.

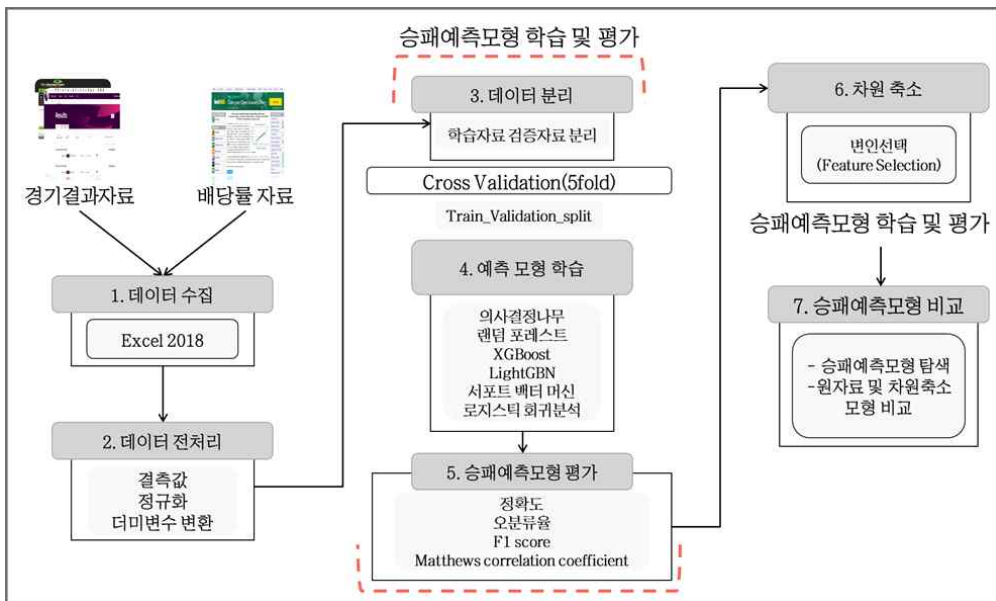


그림 5. 연구절차

첫째, 연구자료들은 Microsoft Excel 2018을 활용하여 수집 및 정리하였다. 둘째, 자료에 대한 전처리 작업을 진행하였다. 구체적으로 결측값을 확인 및 처리하였으며, 변수들의 단위가 상이 하여 정규화 과정을 거쳤다. 또한, 문자로 기록되어 있는 홈팀, 원정팀, 경기결과에 대해서는 수치형 변인으로 변환하였다. 셋째, 전체 자료를 학습자료와 검증자료로 분

리하였다. 모형의 일반화는 어떤 검증자료를 사용하느냐에 따라 크게 달라지기 때문에 다양한 자료로 검증 절차를 반복하는 교차검증은 필수적이며, 이 연구에서는 자료를 5개의 폴더로 나누어 진행하였다. 넷째, 6가지 분류 머신러닝 알고리즘을 활용하여 예측모형을 학습시키고 검증하였다. 다섯째, 정확도, 오분류율, F1-Score, MCC(Matthews Correlation Coefficient)를 이용해서 모형의 타당도를 평가하였다. 여섯째, 차원축소 과정을 실시하였다. 원자료를 활용한 승패예측모형 탐색과정에서 랜덤 포레스트 알고리즘 기반 변인 중요도 값을 산출하여 승패예측모형에 영향을 많이 끼치는 변인 12개를 선택하였다. 마지막은 차원 축소 된 자료를 이용해서 다시 승패예측모형 탐색과정을 진행하였다. 각 단계에 세부 방법들은 자료처리방법에서 자세하게 기술하였다.

2. 연구자료 및 수집방법

잉글랜드 프리미어리그는 시즌제로 구성되어 있으며, 매 시즌은 8월부터 5월까지 진행된다. 1부리그는 20개 팀으로 구성되며, 각 팀은 매 시즌 38경기를 치러 시즌에 총 380경기가 열린다.

이 연구의 첫 번째 연구내용인 경기기록을 활용한 승패예측모형 탐색 및 비교를 위해 잉글랜드 프리미어리그 공식 사이트와 후스코어드닷컴에서 경기기록자료를 수집하였다. 연구의 범위는 09-10시즌부터 18-19시즌으로 선정하였으며, 총 3,800경기에 대한 자료를 수집하였다. 경기결과, 홈팀, 어웨이팀은 모든 연구내용에 포함되는 공통변인으로 설정하였으며, 경기기록 변인의 개수는 총 44개로 설정하였다. <표 1>은 분석에 사용된 경기기록 변인에 대한 설명이다.

두 번째 연구내용인 배당률을 이용한 잉글랜드 프리미어리그 승패예측모형 탐색 및 비교를 위해 football-data에서 배당률 자료를 수집하였다. 결측값이 300개 이상인 변인 15개를 제거하고 분석에서 사용된 배당률 변인의 수는 총 36개이며, <표 2>에 제시하였다. 세 번째 연구내용인 혼합자료를 이용한 잉글랜드 프리미어리그 승패예측모형 탐색 및 비교를

위해 경기기록 자료와 배당률 자료를 통합하였다. 공통변인을 포함한 변인 83개를 활용하여 분석하였다.

표 1. 경기기록 변인

No	경기기록 변인	설명	No	경기기록 변인	설명
1	second time home goal	홈팀 후반전 득점	23	home_cross	홈팀 크로스
2	second time away goal	어웨이팀 후반전 득점	24	away_cross	어웨이팀 크로스
3	half time home goal	홈팀 전반전 득점	25	home_long pass	홈팀 롱패스
4	half time away goal	어웨이팀 전반전 득점	26	away_long pass	어웨이팀 롱패스
5	home team shots	홈팀 슈팅 수	27	home_short pass	홈팀 숏패스
6	away team shots	어웨이팀 슈팅 수	28	away_short pass	어웨이팀 숏패스
7	home team shots on target	홈팀 유효 슈팅 수	29	home_key pass	홈팀 키패스
8	away team shots on target	어웨이팀 유효 슈팅 수	30	away_key pass	어웨이팀 키패스
9	home team pass success	홈팀 패스 성공률	31	home team pass accuracy	홈팀 패스 정확도
10	away team pass success	어웨이팀 패스 성공률	32	away team pass accuracy	어웨이팀 패스 정확도
11	home team dribble won	홈팀 드리블 성공 수	33	home_interception	홈팀 인터셉트
12	away team dribble won	어웨이팀 드리블 성공 수	34	away_interception	어웨이팀 인터셉트
13	home team tackle	홈팀 태클 수	35	home_clearances	홈팀 클리어런스
14	away team tackle	어웨이팀 태클 수	36	away_clearances	어웨이팀 클리어런스
15	home team possession	홈팀 점유율	37	home team foul	홈팀 파울
16	away team possession	어웨이팀 점유율	38	away team foul	어웨이팀 파울
17	home team openplay	홈팀 오픈플레이	39	home team corner	홈팀 코너킥
18	away team openplay	어웨이팀 오픈플레이	40	away team corner	어웨이팀 코너킥
19	home team set piece	홈팀 세트피스	41	home team yellow card	홈팀 경고
20	away team set piece	어웨이팀 세트피스	42	away team yellow card	어웨이팀 경고
21	home_total pass	홈팀 전체 패스	43	home team red card	홈팀 퇴장
22	away_total pass	어웨이팀 전체 패스	44	away team red card	어웨이팀 퇴장

표 2. 배당률 변인

No	배당률 변인	설명
1	B365H	BET365 365 홈 승리 배당률
2	B365D	BET365 366 무승부 배당률
3	B365A	BET365 367 어웨이 승리 배당률
4	BWH	Bet&Win 홈 승리 배당률
5	BWD	Bet&Win 무승부 배당률
6	BWA	Bet&Win 어웨이 승리 배당률
7	IWH	Interwetten 홈 승리 배당률
8	IWD	Interwetten 무승부 배당률
9	IWA	Interwetten 어웨이 승리 배당률
10	PSH	Pinnacle 홈 승리 배당률
11	PSD	Pinnacle 무승부 배당률
12	PSA	Pinnacle 어웨이 승리 배당률
13	WHH	William Hill 홈 승리 배당률
14	WHD	William Hill 무승부 배당률
15	WHA	William Hill 어웨이 승리 배당률
16	VCH	VC BET 홈 승리 배당률
17	VCD	VC BET 무승부 배당률
18	VCA	VC BET 어웨이 승리 배당률
19	BbIX2	Betbrain bookmaker 수
20	BbMxH	Betbrain 홈 승리 최대 배당률
21	BbAvH	Betbrain 홈 승리 평균 배당률
22	BbMxD	Betbrain 무승부 최대 배당률
23	BbAvD	Betbrain 무승부 평균 배당률
24	BbMxA	Betbrain 어웨이 승리 최대 배당률
25	BbAvA	Betbrain 어웨이 승리 평균 배당률
26	BbOU	Betbrain 2.5골 배당 bookmaker 수
27	BbMx>2.5	2.5골 이상 최대 배당률
28	BbAv>2.5	2.5골 이상 평균 배당률
29	BbMx<2.5	2.5골 이하 최대 배당률
30	BbAv<2.5	2.5골 이하 평균 배당률
31	BbAH	Betbrain Asian handicap 배당 bookmaker 수
32	BbAHh	Handicap 크기
33	BbMxAHH	Asian handicap 홈 최대 배당률
34	BbAvAHH	Asian handicap 홈 평균 배당률
35	BbMxAHA	Asian handicap 어웨이 최대 배당률
36	BbAvAHA	Asian handicap 어웨이 평균 배당률

3. 자료처리방법

이 연구는 잉글랜드 프리미어리그의 경기기록과 배당률 자료를 활용한 분류 머신러닝 알고리즘 기반 축구 승패예측모형을 탐색 및 비교하는 것이 목적이다. 연구의 목적을 달성하기 위해 경기기록, 배당률, 혼합자료를 활용하여 축구 승패예측모형을 탐색하였으며, 변수선택 과정을 통한 차원축소 자료를 기반으로 승패예측모형을 탐색 및 비교하여 우수 모형을 확인하였다. 모든 자료처리는 Python 3을 활용하였으며, 자료처리 및 분석에 사용된 Python 라이브러리는 다음 <표 3>과 같다.

표 3. 자료처리 및 분석에 사용된 Python 라이브러리

연구방법	라이브러리
결측값	Pandas
표준화	StandardScaler
범주화 변인 변환	OneHotEncoder, LabelEncoder
교차검증	StratifiedKFold
모형학습	sklearn
모형평가	sklearn.metrics

1) 전처리과정


승패예측모형의 성능을 높이기 위해서는 자료들의 전처리과정은 필수적이다. 이 연구에서 수행한 전처리과정은 과거 N 경기를 활용한 사전변인, 결측값 처리, 표준화, 범주화 변인 변환 단계로 이루어져 있다. 예측모형 전처리과정에서 검증자료의 정보를 학습자료에 포함하는 실수를 종종 범한다. 즉, 결측값을 처리하기 위해서 각 변인의 평균을 구하거나 표준화시키기 위해서 각 변인의 표준편차를 구할 때 학습자료와 검증자료를 분리하지 않은 상태에서 진행한다는 오류를 범한다. 전체 자료를 통한 평균과 표준편차를 사용하면 안 되는 이유는 학습자료에 검증자료들의 정보가 포함되어 있어 예측결과를 신뢰할 수 없기 때문이다. 이 연구에서는 전체 자료를 분리한 후에 학습자료의 평균과 표준편차로 전처리를

진행하였다.

(1) 과거 경기에 따른 사전변인

입력 변인은 결과를 예측하려는 경기의 이전 N 경기 평균값을 사용하였으며, 이 연구에서는 이전 1~5경기에 대해서 모두 실험하여 비교하였다. N 경기의 평균을 구하는 방식은 다음과 같다. 먼저 홈팀과 어웨이팀을 구분하여 모든 시즌에 대한 자료를 통합하였다. 매 시즌 각 팀은 홈과 어웨이를 구분하여 1경기씩 진행하기 때문에 지난 시즌 자료들의 평균값을 사용하였다. 예를 들어 이전 2경기 평균값을 이용하기 위한 전처리과정은 아래 <그림 6>과 같다.

시즌	홈	어웨이	변인1	변인2	결과(승)
09-10	A	B	25	38	Away
10-11	A	B	40	35	Home
11-12	A	B	28	32	Away



이전 2경기 평균

시즌	홈	어웨이	변인1	변인2	결과(승)
11-12	A	B	32.5	36.5	Away

그림 6. 사전변인 전처리과정

(2) 결측값 처리

수집한 자료에서 나타난 결측값은 모두 해당 자료를 나타냈다. 처음 51개의 해당 자료를 수집하였을 때 30% 이상 결측값이 존재하는 변인 15개를 발견하였다. 결측값이 존재하는 사례를 제거하기에는 자료가 왜곡될 가능성이 있어 해당 변인만 삭제하였다(박재현, 강민수, 이진오, 강상조, 2005). 하지만 <표 4>에 제시한 변인들은 전체 사례 수의 1% 미만이기 때문에 각 변인의 평균으로 대체하였다.

표 4. 결측값 처리 변인

변인	결측값
BWA	1
BWD	1
BWH	1
BbAH	10
BbAHh	10
BbAvAHA	10
BbAvAHH	10
BbMxAHA	10
BbMxAHH	10
IWA	1
IWD	1
IWH	1

(3) 표준화


이 연구에서 사용된 변인들의 단위는 다르다. 구체적으로 경고와 파울은 단순 빈도이지만 패스 정확도는 비율로 측정되었기 때문에 표준화는 필수적이다. 표준화는 자료들을 동일한 범위의 값으로 만들어 학습속도를 증가시키며, 각 변인이 모형을 학습시키는 데 있어서 비슷한 기여도를 가질 수 있도록 한다(조준모, 2019). 이 연구에서는 Z 점수를 이용하여 표준화 작업을 진행하였다. Z 점수는 원점수의 상대적 위치를 알려주는 점수를 말한다. 분포가 정규분포라는 가정에서 평균을 0으로 표준편차를 1로 하는 점수이다(강상조, 1994). 표준화 작업을 위해 Python의 StandardScaler 라이브러리를 사용하였다.

(4) 범주형 변인 변환

수집된 자료 중 경기결과 변인은 홈팀 승리는 H, 어웨이팀 승리는 A, 무승부는 D로 표기되어 있으며 홈팀과 어웨이팀은 각 팀의 이름 그대로 수집하였다. 이러한 텍스트 형태로 되어있는 범주형 자료는 수치형으로 변경하였을 때 머신러닝 알고리즘이 더 잘 학습할 수 있다(Lin, Wu, Liu, Xia, & Bhattarai, 2020). 이 연구에서는 Python에서 제공하는

LabelEncoder 방법을 사용하였다. 각 범주에 숫자를 부여하는 방식을 말하는데 H(홈 승리)는 2, A(어웨이팀 승리)는 0, D(무승부)는 1을 부여하는 방식이다. 홈팀과 어웨이팀 자료의 수치형 변환은 OneHotEncoder 방식을 사용하여 아래 <그림 7>과 같이 변경하였다. 경기결과에 사용한 LabelEncoder의 단점은 범주의 종류가 많을 때, 부여된 숫자들이 서열의 성질을 가지게 되어 큰 숫자가 부여된 범주는 더 높은 가중치를 가지게 된다는 것이다 (Raheel, 2018).

ID	Home		Away	
1	A		B	
2	C		D	
3	E		F	



ID	A	B	C	D	E	F
1	1	1	0	0	0	0
2	0	0	1	1	0	0
3	0	0	0	0	1	1

그림 7. OneHotEncoder 방법

2) 교차검증(Cross Validation)

분류 머신러닝 알고리즘을 활용해서 모델을 개발할 때 모델을 학습시키기 위한 학습자료와 모형의 성능을 평가하기 위한 검증자료가 필요하다. 검증자료를 통해서 모형을 평가하였더라도 실제 상황에서 바로 사용하기에는 어려움이 있어 여러 번의 검증 절차를 거치는 교차검증이 필요하다(Browne, 2000). 교차검증은 학습자료를 K개의 폴더로 나눈다. 일반적으로 5개의 폴더 또는 10개의 폴더로 나누며(송상윤, 2015) 이 연구에서는 5개의 폴더로

설정하였다. 5개의 폴더 중 4개는 학습자료로 사용하고 나머지 1개는 검증자료로 설정하였다. 이 과정을 5개의 폴더가 교차하면서 5번 시행하였으며 각 시행에 대한 정확도를 확인하였다. 최적의 모형을 구성하기 위해 5개의 예측모형을 합쳐 하나의 최종 모형으로 만들었다.

3) 모형평가

최적의 승패예측모형을 판단하기 위해 모형평가는 필수적이다. 모형을 평가하는 방법으로 각 모형에 대한 타당도 지수를 근거로 하였다. 구체적으로 교차검증을 통해서 최종 모형을 선택한 후에 검증자료를 정확하게 분류하는지에 대한 타당도 지수를 살펴보았다. 다른 스포츠에 대한 타당도 지수를 산출할 때는 승리, 패배에 대한 분류를 기반으로 한 이원 분류표를 활용해서 검증한다. 그러나 축구 경기결과는 승리, 패배, 무승부로 명확하게 구분되기 때문에 삼원 분류표를 이용하여 모형의 타당도 지수를 산출하였다.

삼원 분류표를 이용한 타당도 지수는 분류정확도(Accuracy), 오분류율(Error rate), F-1 score, Matthews correlation coefficient(MCC)로 설정하였다. 분류정확도와 오분류율은 분류모형을 평가하는 방법 중 가장 일반적으로 사용되는 지수이다(Tharwat, 2020). 정밀도와 재현율의 조화평균 지수인 F-1 score과 실제 분류와 예측모형 분류 사이의 상관계수를 나타내는 Matthews correlation coefficient(MCC)는 불균형 분류 문제를 평가하기 위한 타당도 지수로 평가받고 있다. 각 타당도 지수를 산출하기 위해서는 홈팀 승리, 어웨이팀 승리, 무승부의 삼원 분류표를 산출해야 하지만 이 연구에서는 Python의 내장되어있는 함수를 이용하여 타당도 지수를 산출하였으며, 삼원 분류표의 예시와 공식은 아래와 같다.

		실제 경기결과		
		어웨이	무승부	홈
승패예측모형	어웨이	TP (True Positive)	FP (False Positive)	FP (False Positive)
	무승부	FN (False Negative)	TN (True Negative)	TN (True Negative)
	홈	FN (False Negative)	TN (True Negative)	TN (True Negative)

그림 8. 어웨이팀 삼원 분류표 예시

<그림 8>은 어웨이팀을 기준으로 나타낸 삼원 분류표 예시이다. 만약 승패예측모형이 어웨이팀 승리로 예측하고 실제로 어웨이팀이 승리했다면 True Positive(TP)로 분류한다. True Negative(TN)는 승패예측모형에서 어웨이팀 승리가 아니라고 예측되고 실제 결과에서도 어웨이팀 승리가 아닌 무승부나 홈팀의 승리인 경우이다. False Positive(FP)는 승패예측모형에서 어웨이팀 승리로 예측되었지만 실제로 어웨이팀 승리가 아닌 경우를 나타낸다. False Negative(FN)는 어웨이팀이 실제로 경기에서 승리하였지만, 승패예측모형은 어웨이팀이 패배 또는 무승부라고 예측한 경우를 의미한다.

(1) 분류정확도(Accuracy)

분류정확도는 가장 빈번하게 사용되는 타당도 지수이며, 예측한 결과 중 정확하게 분류한 비율을 나타낸다. 수식은 아래 <수식 6>과 같다.

$$\text{〈수식 6〉}$$

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

(2) 오분류율(Error Rate)

오분류율은 분류정확도의 반대되는 개념으로써 예측한 결과 중 부정확하게 예측한 비율을 나타낸다. 수식은 아래 <수식 7>과 같다.

<수식 7>

$$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN} = 1 - Accuracy$$

(3) F-1 score

F-1 score은 분류의 클래스 즉, 홈팀 승, 어웨이팀 승, 무승부의 자료가 일정한 비율로 존재하지 않고 불균형할 때 사용하는 지수이며, 정밀도와 재현률의 조화평균을 의미한다. 수식은 아래 <수식 8>과 같다.

<수식 8>

$$F1 - score = 2 \times \frac{(TP / (TP + FP) \times TP / (TP + FN))}{(TP / (TP + FP) + TP / (TP + FN))}$$

(4) Matthews correlation coefficient(MCC)

Matthews correlation coefficient는 예측모형을 이용한 분류결과와 실제 경기결과와의 상관계수를 의미하며, F-1 score과 함께 불균형 자료에 적합한 타당도이다. 구체적인 수식은 아래 <수식 9>와 같다.

<수식 9>

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

4) 차원 축소(Dimensionality Reduction)

차원축소란 불필요한 변수를 제거하여 모형을 단순화하는 방법이며, 1) 계산의 복잡성 감소 2) 모형 일반화 3) 자료 저장 비용 감소의 목적이 있다(Tang, Alelyani, & Liu, 2014). 차원축소는 변수추출과 변수선택으로 나누어질 수 있으며 변수추출은 종속변수에 영향을

많이 미치는 변인들을 결합해 새로운 변인들을 만드는 방법이다. 대표적으로 주성분분석(principle component analysis: PCA)과 선형판별분석(linear discriminant analysis: LDA)을 예로 들 수 있다. 변인선택은 주어진 자료 중 종속변인의 범주들에 가장 큰 영향을 미치는 변인의 부분집합을 찾는 방법이며, filter, wrapper 그리고 embedded 방법으로 구분된다(정인범, 2018).

이 연구에서는 변인선택에 변인 중요도(feature importance)값을 산출하는 MDG(Mean Decrease Gini) 방법을 사용하여 차원축소 하였다. 변인추출은 새롭게 형성된 변인들이 기존에 어떤 변인들로 인해서 형성되었는지를 파악하기 어렵다는 단점 때문에 사용하지 않았다. MDG는 트리 알고리즘 모형에서 변인의 중요도를 측정하는 값이며, 한 변인이 모형에 적용되었을 때 전체적으로 분류의 불순도(impurity)가 얼마나 감소하는지를 의미한다.

불순도는 지니지수(Gini Index)로 측정된다(Zhang, & Ma, 2012). 지니지수는 트리 구조에서 한 노드의 불순도를 측정한 값이다. 부모 노드에 한 변인을 기준으로 분류해서 자식 노드가 나오면 부모노드의 지니지수와 자식 노드의 지니지수를 각각 산출할 수 있다. 분류가 잘 이루어졌을 경우, 부모 노드의 지니지수에 비해 분류 후 자식 노드의 지니지수가 감소할 것이며, 이는 해당 변인에서 지니지수의 감소 정도(decrease of gini index)가 커진다는 것을 의미한다. 즉, 각 트리에서 해당 변인을 기준으로 분류하는 지점에서 불순도 감소 정도의 총합을 계산한 후, 모든 트리의 값들을 평균낸 것이 MDG이다(RColorBrewer, & Liaw, 2018).

IV. 연구결과

이 연구의 목적은 잉글랜드 프리미어리그 과거 경기기록과 배당률 자료를 활용하여 분류 머신러닝 알고리즘 기반 축구 승패예측모형을 탐색하는 것이다. 연구의 목적을 달성하기 위하여 이 연구에서는 다음 세 가지의 연구내용을 제시하였다. 첫째, 과거 경기기록을 활용한 분류 머신러닝 알고리즘 기반 승패예측모형을 탐색하고 비교하였다. 둘째, 배당률 자료를 이용한 분류 머신러닝 알고리즘 기반 승패예측모형을 탐색하고 비교하였으며, 셋째, 경기기록과 배당률을 모두 결합한 혼합자료를 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 및 탐색 및 비교하였다.

각 연구내용의 세부내용은 세 단계로 나누어져 있다. 첫 번째 단계는 원자료(raw data)를 사용한 승패예측모형을 탐색하였다. 두 번째 단계는 차원축소 과정을 거친 자료를 활용해서 승패예측모형을 탐색하였으며, 마지막으로 이전 단계에서 구축한 모형들을 비교하여 최적 모형을 확인하였다.

1. 경기기록을 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교

1) 경기기록을 활용한 승패예측모형 탐색

6가지 분류 머신러닝 알고리즘을 사용하여 잉글랜드 프리미어리그 승패예측모형을 탐색하였다. 사용된 입력 변인은 과거 경기기록에 관한 자료를 투입하였으며, 이전 1경기 자료부터 5경기 자료까지의 조합을 모두 분석하였다. 각 분류 알고리즘에 따라 어떤 조합의 이전 경기기록 자료가 적합한지 알아보기 위하여 타당도 검증결과와 순위를 산출하였으며, 결과는 다음과 같다.

(1) 경기기록에 따른 의사결정나무 승패예측모형 타당도

의사결정나무 알고리즘을 기반으로 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 4경기 자료의 평균을 이용한 모형이 ACC(0.473), ERR(0.527), F-1(0.469), MCC(0.166)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 자료의 평균을 이용한 모형이 ACC(0.454), ERR(0.546), F-1(0.449), MCC(0.146)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 5>와 같다.

표 5. 경기기록에 따른 의사결정나무 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.447	3	0.553	3	0.447	3	0.147	2	3
2경기전	0.454	2	0.546	2	0.449	2	0.146	3	2
3경기전	0.442	4	0.558	4	0.443	4	0.141	4	4
4경기전	0.473	1	0.527	1	0.469	1	0.166	1	1
5경기전	0.418	5	0.582	5	0.414	5	0.084	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(2) 경기기록에 따른 랜덤 포레스트 승패예측모형 타당도

랜덤 포레스트 알고리즘을 기반으로 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 자료의 평균을 이용한 모형이 ACC(0.532), ERR(0.468), F-1(0.452), MCC(0.240)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 자료의 평균을 이용한 모형이 ACC(0.531), ERR(0.469), F-1(0.449), MCC(0.225)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 6>과 같다.

표 6. 경기기록에 따른 랜덤 포레스트 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.516	4	0.484	4	0.438	4	0.197	4	4
2경기전	0.531	2	0.469	2	0.449	3	0.225	2	2
3경기전	0.532	1	0.468	1	0.452	1	0.240	1	1
4경기전	0.531	2	0.469	2	0.451	2	0.215	3	3
5경기전	0.456	5	0.544	5	0.399	5	0.106	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(3) 경기기록에 따른 XGBoost 승패예측모형 타당도

XGBoost 알고리즘을 기반으로 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 2경기 자료의 평균을 이용한 모형이 ACC(0.539), ERR(0.462), F-1(0.457), MCC(0.240)로 가장 높은 순위를 나타냈다. 두 번째는 과거 4경기 자료의 평균을 이용한 모형이 ACC(0.523), ERR(0.477), F-1(0.452), MCC(0.201)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 7>과 같다.

표 7. 경기기록에 따른 XGBoost 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.511	4	0.489	4	0.440	5	0.191	4	4
2경기전	0.539	1	0.462	1	0.457	1	0.240	1	1
3경기전	0.522	3	0.478	3	0.449	3	0.223	2	3
4경기전	0.523	2	0.477	2	0.452	2	0.201	3	2
5경기전	0.478	5	0.522	5	0.446	4	0.145	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(4) 경기기록에 따른 LightGBM 승패예측모형 타당도

LightGBM 알고리즘을 기반으로 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 자료의 평균을 이용한 모형이 ACC(0.519), ERR(0.481), F-1(0.434), MCC(0.209)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 자료의 평균을 이용한 모형이 ACC(0.515), ERR(0.485), F-1(0.418), MCC(0.187)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 8>과 같다.

표 8. 경기기록에 따른 LightGBM 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.489	4	0.511	4	0.408	4	0.131	5	5
2경기전	0.515	2	0.485	2	0.418	3	0.187	2	2
3경기전	0.519	1	0.481	1	0.434	1	0.209	1	1
4경기전	0.504	3	0.496	3	0.407	5	0.137	3	3
5경기전	0.484	5	0.517	5	0.427	2	0.137	3	4

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(5) 경기기록에 따른 로지스틱 회귀분석 승패예측모형 타당도

로지스틱 회귀분석 알고리즘을 기반으로 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 1경기 자료를 이용한 모형이 ACC(0.520), ERR(0.480), F-1(0.485), MCC(0.217)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 자료의 평균을 이용한 모형이 ACC(0.513), ERR(0.487), F-1(0.483), MCC(0.204)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 9>와 같다.

표 9. 경기기록에 따른 로지스틱 회귀분석 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.520	1	0.480	1	0.485	1	0.217	1	1
2경기전	0.513	2	0.487	2	0.483	2	0.204	2	2
3경기전	0.499	3	0.501	3	0.461	3	0.187	3	3
4경기전	0.492	4	0.508	4	0.459	4	0.158	4	4
5경기전	0.462	5	0.539	5	0.452	5	0.131	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(6) 경기기록에 따른 서포트 벡터 머신 승패예측모형 타당도

서포트 벡터 머신 알고리즘을 기반으로 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 2경기 자료의 평균을 이용한 모형이 ACC(0.526), ERR(0.474), F-1(0.433), MCC(0.212)로 가장 높은 순위를 나타냈다. 두 번째는 과거 1경기 자료의 평균을 이용한 모형이 ACC(0.518), ERR(0.482), F-1(0.437), MCC(0.196)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 10>과 같다.

표 10. 경기기록에 따른 서포트 벡터 머신 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.518	2	0.482	2	0.437	1	0.196	3	2
2경기전	0.526	1	0.474	1	0.433	3	0.212	1	1
3경기전	0.517	3	0.483	3	0.436	2	0.207	2	3
4경기전	0.516	4	0.485	4	0.428	5	0.171	5	4
5경기전	0.506	5	0.495	5	0.432	4	0.178	4	4

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(7) 경기기록 기반 최적 승패예측모형 선택

과거 경기기록 조합(1경기 전~5경기 전 평균)에 따른 6가지 분류 머신러닝 알고리즘의 승패예측모형 타당도를 검증하였다. 각 알고리즘의 최적 승패예측모형은 <표 11>에 나타냈으며, 최적 승패예측모형 간의 순위를 비교하였다. 그 결과 XGBoost 알고리즘을 이용한 승패예측모형(M3)이 ACC(0.539), ERR(0.462), F-1(0.457), MCC(0.240)로 가장 높은 순위를 나타냈다. 두 번째는 랜덤 포레스트 알고리즘을 이용한 승패예측모형(M2)이 ACC(0.532), ERR(0.468), F-1(0.452), MCC(0.240)로 높은 순위를 기록하였다.

표 11. 경기기록에 따른 최적 승패예측모형 타당도 비교

모형	과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
M1	4경기전	0.473	6	0.527	6	0.469	2	0.166	6	5
M2	3경기전	0.532	2	0.468	2	0.452	4	0.240	2	2
M3	2경기전	0.539	1	0.462	1	0.457	3	0.240	1	1
M4	3경기전	0.519	5	0.481	5	0.434	5	0.209	5	5
M5	1경기전	0.520	4	0.480	4	0.485	1	0.217	3	3
M6	2경기전	0.526	3	0.474	3	0.433	6	0.212	4	4

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

· M1: 의사결정나무, · M2: 랜덤 포레스트, · M3: XGBoost, · M4: LightGBM, · M5: 로지스틱 회귀분석, · M6: 서포트벡터머신

2) 차원축소 경기기록을 활용한 승패예측모형 탐색

연구내용 1-2에서는 차원축소 과정을 통해 변인이 축소된 경기기록 자료를 활용한 승패 예측모형을 탐색하였다. 변인 중요도에 따른 선택된 변인은 상위 12개로 한정하였으며, 자료의 조합에 따른 변인 항목은 부록에 첨부하였다. 각 분류 머신러닝 알고리즘의 결과는 다음과 같다.

(1) 차원축소 과거 경기기록에 따른 의사결정나무 승패예측모형 타당도

의사결정나무 알고리즘을 기반으로 차원축소 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 자료의 평균을 이용한 모형이 ACC(0.468), ERR(0.532), F-1(0.462), MCC(0.167)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 자료의 평균을 이용한 모형이 ACC(0.458), ERR(0.542), F-1(0.460), MCC(0.161)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 12>와 같다.

표 12. 차원축소 경기기록에 따른 의사결정나무 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.424	3	0.576	3	0.423	3	0.102	3	3
2경기전	0.458	2	0.542	2	0.460	2	0.161	2	2
3경기전	0.468	1	0.532	1	0.462	1	0.167	1	1
4경기전	0.419	4	0.581	4	0.417	4	0.087	4	4
5경기전	0.368	5	0.632	5	0.372	5	0.019	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(2) 차원축소 과거 경기기록에 따른 랜덤 포레스트 승패예측모형 타당도

랜덤 포레스트 알고리즘을 기반으로 차원축소 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 자료의 평균을 이용한 모형이 ACC(0.548), ERR(0.452), F-1(0.470), MCC(0.270)로 가장 높은 순위를 나타냈다. 두 번째는 과거 4경기 자료의 평균을 이용한 모형이 ACC(0.547), ERR(0.454), F-1(0.466), MCC(0.245)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 13>과 같다.

표 13. 차원축소 경기기록에 따른 랜덤 포레스트 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.514	4	0.486	4	0.437	4	0.194	4	4
2경기전	0.539	3	0.462	3	0.458	3	0.240	3	3
3경기전	0.548	1	0.452	1	0.470	1	0.270	1	1
4경기전	0.547	2	0.454	2	0.466	2	0.245	2	2
5경기전	0.462	5	0.539	5	0.419	5	0.119	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(3) 차원축소 과거 경기기록에 따른 XGBoost 승패예측모형 타당도

XGBoost 알고리즘을 기반으로 차원축소 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 2경기 자료의 평균을 이용한 모형이 ACC(0.539), ERR(0.462), F-1(0.457), MCC(0.240)로 가장 높은 순위를 나타냈다. 두 번째는 과거 4경기 자료의 평균을 이용한 모형이 ACC(0.522), ERR(0.478), F-1(0.452), MCC(0.201)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 14>와 같다.

표 14. 차원축소 경기기록에 따른 XGBoost 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.511	4	0.489	4	0.440	5	0.191	4	4
2경기전	0.539	1	0.462	1	0.457	1	0.240	1	1
3경기전	0.522	3	0.478	3	0.449	3	0.223	2	3
4경기전	0.522	2	0.478	2	0.452	2	0.201	3	2
5경기전	0.478	5	0.522	5	0.446	4	0.145	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(4) 차원축소 과거 경기기록에 따른 LightGBM 승패예측모형 타당도

LightGBM 알고리즘을 기반으로 차원축소 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 자료의 평균을 이용한 모형이 ACC(0.519), ERR(0.481), F-1(0.434), MCC(0.209)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 자료의 평균을 이용한 모형이 ACC(0.515), ERR(0.485), F-1(0.418), MCC(0.187)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 15>와 같다.

표 15. 차원축소 경기기록에 따른 LightGBM 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.489	4	0.511	4	0.408	4	0.131	5	5
2경기전	0.515	2	0.485	2	0.418	3	0.187	2	2
3경기전	0.519	1	0.481	1	0.434	1	0.209	1	1
4경기전	0.504	3	0.496	3	0.407	5	0.137	4	3
5경기전	0.484	5	0.517	5	0.427	2	0.137	3	3

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(5) 차원축소 과거 경기기록에 따른 로지스틱 회귀분석 승패예측모형 타당도

로지스틱 회귀분석 알고리즘을 기반으로 차원축소 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 4경기 자료의 평균을 이용한 모형이 ACC(0.547), ERR(0.454), F-1(0.509), MCC(0.255)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 자료의 평균을 이용한 모형이 ACC(0.553), ERR(0.447), F-1(0.487), MCC(0.220)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 16>과 같다.

표 16. 차원축소 경기기록에 따른 로지스틱 회귀분석 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.524	3	0.476	4	0.484	3	0.222	2	3
2경기전	0.553	1	0.447	1	0.487	2	0.220	3	2
3경기전	0.509	4	0.491	5	0.462	5	0.201	4	5
4경기전	0.547	2	0.454	2	0.509	1	0.255	1	1
5경기전	0.489	5	0.475	3	0.470	4	0.168	5	4

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(6) 차원축소 과거 경기기록에 따른 서포트 벡터 머신 승패예측모형 타당도

서포트 벡터 머신 알고리즘을 기반으로 차원축소 과거 경기기록 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 4경기 자료의 평균을 이용한 모형이 ACC(0.539), ERR(0.461), F-1(0.454), MCC(0.225)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 자료의 평균을 이용한 모형이 ACC(0.531), ERR(0.469), F-1(0.441), MCC(0.224)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전)기록에 따른 타당도 검증결과는 다음 <표 17>과 같다.

표 17. 차원축소 경기기록에 따른 서포트 벡터 머신 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.520	4	0.480	4	0.440	3	0.203	4	4
2경기전	0.531	2	0.469	2	0.441	2	0.224	2	2
3경기전	0.522	3	0.478	3	0.440	4	0.216	3	3
4경기전	0.539	1	0.461	1	0.454	1	0.225	1	1
5경기전	0.478	5	0.522	5	0.415	5	0.133	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(7) 차원축소 경기기록 자료 기반 최적 승패예측모형 선택

차원축소 과정을 거친 과거 경기자료 조합(1경기 전~5경기 전 평균)에 따른 6가지 분류 머신러닝 알고리즘의 승패예측모형 타당도를 검증하였다. 각 알고리즘의 최적 승패예측모형은 <표 18>에 나타냈으며, 최적 승패예측모형 간의 순위를 비교하였다. 그 결과 랜덤 포레스트 알고리즘을 이용한 승패예측모형(RM2)이 ACC(0.548), ERR(0.452), F-1(0.470), MCC(0.270)로 가장 높은 순위를 나타냈다. 두 번째는 로지스틱 회귀분석 알고리즘을 이용한 승패예측모형(RM5)이 ACC(0.547), ERR(0.454), F-1(0.509), MCC(0.255)로 높은 순위를 기록하였다.

표 18. 차원축소 경기기록에 따른 최적 승패예측모형 타당도 비교

모형	과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
RM1	3경기전	0.468	6	0.532	6	0.462	3	0.167	6	5
RM2	3경기전	0.548	1	0.452	1	0.470	2	0.270	1	1
RM3	2경기전	0.539	4	0.462	4	0.457	4	0.240	3	3
RM4	3경기전	0.519	5	0.481	5	0.434	6	0.209	5	5
RM5	4경기전	0.547	2	0.454	2	0.509	1	0.255	2	2
RM6	4경기전	0.539	3	0.461	3	0.454	5	0.225	4	3

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

· RM(Reduction Model)1: 차원축소 의사결정나무, · RM(Reduction Model)2: 차원축소 랜덤 포레스트,

· RM(Reduction Model)3: 차원축소 XGBoost, · RM(Reduction Model)4: 차원축소 LightGBM,

· RM(Reduction Model)5: 차원축소 로지스틱 회귀분석, · RM(Reduction Model)6: 차원축소 서포트 벡터 머신

3) 경기기록 승패예측모형과 차원축소 경기기록 승패예측모형 비교

경기기록 승패예측모형과 차원축소 경기기록 승패예측모형 간의 순위 비교 결과는 다음 <표 19>와 같다. 차원축소 경기기록 자료와 랜덤 포레스트 알고리즘을 함께 사용한 승패예측모형(RM2)이 ACC(0.548), ERR(0.452), F-1(0.470), MCC(0.270)로 가장 높은 순위로 나타났다. 1위부터 5위까지의 모형은 RM2> RM5> RM3, M3 >RM6으로 나타났으며, 차원축소 자료를 사용한 승패예측모형이 기존 자료를 사용한 모형보다 타당도가 높거나 비슷한 수준을 유지하였다.

표 19. 최적 승패예측모형 종합 타당도 및 순위 비교

모형	과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
M1	4경기전	0.473	11	0.527	11	0.469	4	0.166	12	11
M2	3경기전	0.532	6	0.468	6	0.452	9	0.240	5	7
M3	2경기전	0.539	4	0.462	4	0.457	6	0.240	3	3
M4	3경기전	0.519	9	0.481	9	0.434	10	0.209	9	9
M5	1경기전	0.520	8	0.480	8	0.485	2	0.217	7	6
M6	2경기전	0.526	7	0.474	7	0.433	12	0.212	8	8
RM1	3경기전	0.468	12	0.532	12	0.462	5	0.167	11	12
RM2	3경기전	0.548	1	0.452	1	0.470	3	0.270	1	1
RM3	2경기전	0.539	4	0.462	4	0.457	6	0.240	3	3
RM4	3경기전	0.519	9	0.481	9	0.434	10	0.209	9	9
RM5	4경기전	0.547	2	0.454	2	0.509	1	0.255	2	2
RM6	4경기전	0.539	3	0.461	3	0.454	8	0.225	6	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

· M1: 의사결정나무, · M2: 랜덤 포레스트, · M3: XGBoost, · M4: LightGBM, · M5: 로지스틱 회귀분석, · M6: 서포트벡터머신

· RM(Reduction Model)1: 차원축소 의사결정나무, · RM(Reduction Model)2: 차원축소 랜덤 포레스트,

· RM(Reduction Model)3: 차원축소 XGBoost, · RM(Reduction Model)4: 차원축소 LightGBM,

· RM(Reduction Model)5: 차원축소 로지스틱 회귀분석, · RM(Reduction Model)6: 차원축소 서포트 벡터 머신

2 배당률을 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교

1) 배당률을 활용한 승패예측모형 탐색

6가지 분류 머신러닝 알고리즘을 사용하여 잉글랜드 프리미어리그 승패예측모형을 탐색하였다. 사용된 입력 변인은 배당률 자료를 투입하였으며, 이전 1경기 자료부터 5경기 자료까지의 조합을 모두 분석하였다. 각 분류 알고리즘에 따라 어떤 조합의 배당률 자료가 적합한지 알아보기 위하여 타당도 검증결과와 순위를 산출하였으며, 결과는 다음과 같다.

(1) 배당률에 따른 의사결정나무 승패예측모형 타당도

의사결정나무 알고리즘을 기반으로 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 배당률의 평균을 이용한 모형이 ACC(0.496), ERR(0.504), F-1(0.494), MCC(0.217)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 배당률의 평균을 이용한 모형이 ACC(0.485), ERR(0.515), F-1(0.483), MCC(0.195)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 20>과 같다.

표 20. 배당률에 따른 의사결정나무 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.449	3	0.552	3	0.449	3	0.148	3	3
2경기전	0.485	2	0.515	2	0.483	2	0.195	2	2
3경기전	0.496	1	0.504	1	0.494	1	0.217	1	1
4경기전	0.423	5	0.578	5	0.430	5	0.110	5	5
5경기전	0.440	4	0.560	4	0.441	4	0.123	4	4

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(2) 배당률에 따른 랜덤 포레스트 승패예측모형 타당도

랜덤 포레스트 알고리즘을 기반으로 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기과 4경기 배당률의 평균을 이용한 모형들이 가장 높은 순위를 나타냈다. 과거 3경기 배당률의 평균을 이용한 모형은 ACC(0.558), ERR(0.442), F-1(0.528), MCC(0.289)의 타당도 지수가 나타났으며, 과거 4경기 배당률의 평균을 이용한 모형이 ACC(0.558), ERR(0.442), F-1(0.530), MCC(0.271)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 21>과 같다.

표 21. 배당률에 따른 랜덤 포레스트 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.535	5	0.465	5	0.487	5	0.237	5	5
2경기전	0.539	4	0.462	4	0.499	4	0.245	4	4
3경기전	0.558	1	0.442	1	0.528	2	0.289	1	1
4경기전	0.558	1	0.442	1	0.530	1	0.271	2	1
5경기전	0.550	3	0.451	3	0.500	3	0.257	3	3

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(3) 배당률에 따른 XGBoost 승패예측모형 타당도

XGBoost 알고리즘을 기반으로 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 1경기 배당률을 이용한 모형이 ACC(0.563), ERR(0.437), F-1(0.495), MCC(0.286)로 가장 높은 순위를 나타냈다. 두 번째는 과거 4경기 배당률의 평균을 이용한 모형이 ACC(0.554), ERR(0.446), F-1(0.520), MCC(0.266)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 22>와 같다.

표 22. 배당률에 따른 XGBoost 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.563	1	0.437	1	0.495	3	0.286	1	1
2경기전	0.535	4	0.465	4	0.479	5	0.235	4	4
3경기전	0.553	3	0.447	3	0.517	2	0.279	2	3
4경기전	0.554	2	0.446	2	0.520	1	0.266	3	2
5경기전	0.522	5	0.478	5	0.495	3	0.222	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(4) 배당률에 따른 LightGBM 승패예측모형 타당도

LightGBM 알고리즘을 기반으로 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 배당률의 평균을 이용한 모형이 ACC(0.550), ERR(0.450), F-1(0.493), MCC(0.270)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 배당률의 평균을 이용한 모형이 ACC(0.544), ERR(0.456), F-1(0.486), MCC(0.250)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 23>과 같다.

표 23. 배당률에 따른 LightGBM 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.533	4	0.467	4	0.466	5	0.226	4	4
2경기전	0.544	2	0.456	2	0.486	2	0.250	2	2
3경기전	0.550	1	0.450	1	0.493	1	0.270	1	1
4경기전	0.527	5	0.473	5	0.472	3	0.202	5	5
5경기전	0.539	3	0.462	3	0.470	4	0.229	3	3

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(5) 배당률에 따른 로지스틱 회귀분석 승패예측모형 타당도

로지스틱 회귀분석 알고리즘을 기반으로 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 4경기 배당률의 평균을 이용한 모형이 ACC(0.566), ERR(0.434), F-1(0.539), MCC(0.289)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 배당률의 평균을 이용한 모형이 ACC(0.524), ERR(0.476), F-1(0.492), MCC(0.225)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 24>와 같다.

표 24. 배당률에 따른 로지스틱 회귀분석 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.516	4	0.484	4	0.480	5	0.207	4	4
2경기전	0.524	2	0.476	2	0.492	2	0.225	3	2
3경기전	0.519	3	0.481	3	0.490	3	0.226	2	3
4경기전	0.566	1	0.434	1	0.539	1	0.289	1	1
5경기전	0.500	5	0.500	5	0.486	4	0.193	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(6) 배당률에 따른 서포트 벡터 머신 승패예측모형 타당도

서포트 벡터 머신 알고리즘을 기반으로 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 4경기 배당률의 평균을 이용한 모형이 ACC(0.574), ERR(0.426), F-1(0.481), MCC(0.298)로 가장 높은 순위를 나타냈다. 두 번째는 과거 1경기 배당률을 이용한 모형이 ACC(0.550), ERR(0.450), F-1(0.467), MCC(0.260)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 25>와 같다.

표 25. 배당률에 따른 서포트 벡터 머신 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.550	2	0.450	2	0.467	2	0.260	3	2
2경기전	0.542	5	0.458	5	0.451	5	0.248	5	5
3경기전	0.545	3	0.455	3	0.458	3	0.262	2	3
4경기전	0.574	1	0.426	1	0.481	1	0.298	1	1
5경기전	0.544	4	0.456	4	0.454	4	0.251	4	4

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(7) 배당률 기반 최적 승패예측모형 선택

배당률 자료 조합(1경기 전~5경기 전 평균)에 따른 6가지 분류 머신러닝 알고리즘의 승패예측모형 타당도를 검증하였다. 각 알고리즘의 최적 승패예측모형은 <표 26>에 나타냈으며, 최적 승패예측모형 간의 순위를 비교하였다. 그 결과 로지스틱 회귀분석 알고리즘을 이용한 승패예측모형(M5)이 ACC(0.566), ERR(0.434), F-1(0.539), MCC(0.289)로 가장 높은 순위를 나타냈다. 두 번째는 서포트 벡터 머신 알고리즘을 이용한 승패예측모형(M6)이 ACC(0.574), ERR(0.426), F-1(0.481), MCC(0.298)로 높은 순위를 기록하였다.

표 26. 배당률에 따른 최적 승패예측모형 타당도 비교

모형	과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
M1	3경기전	0.496	6	0.504	6	0.494	4	0.217	6	6
M2	3경기전	0.558	4	0.442	4	0.528	2	0.289	2	3
M3	1경기전	0.563	3	0.437	3	0.495	3	0.286	4	4
M4	3경기전	0.550	5	0.450	5	0.493	5	0.270	5	5
M5	4경기전	0.566	2	0.434	2	0.539	1	0.289	3	1
M6	4경기전	0.574	1	0.426	1	0.481	6	0.298	1	2

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

· M1: 의사결정나무, · M2: 랜덤 포레스트, · M3: XGBoost, · M4: LightGBM, · M5: 로지스틱 회귀분석, · M6: 서포트벡터머신

2) 차원축소 배당률을 활용한 승패예측모형 탐색

연구내용 2-2에서는 차원축소 과정을 통해 변인이 축소된 배당률 자료를 활용한 승패예측모형을 탐색하였다. 변인 중요도에 따른 선택된 변인은 상위 12개로 한정하였으며, 자료의 조합에 따른 변인 항목은 부록에 첨부하였다. 각 분류 머신러닝 알고리즘의 결과는 다음과 같다.

(1) 차원축소 배당률에 따른 의사결정나무 승패예측모형 타당도

의사결정나무 알고리즘을 기반으로 차원축소 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 5경기 배당률의 평균을 이용한 모형이 ACC(0.467), ERR(0.533), F-1(0.469), MCC(0.168)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 배당률을 이용한 모형이 ACC(0.458), ERR(0.542), F-1(0.454), MCC(0.146)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 27>과 같다.

표 27. 차원축소 배당률에 따른 의사결정나무 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.447	4	0.553	4	0.448	4	0.142	4	4
2경기전	0.458	2	0.542	2	0.454	2	0.146	2	2
3경기전	0.444	5	0.556	5	0.445	5	0.139	5	5
4경기전	0.450	3	0.550	3	0.453	3	0.142	3	3
5경기전	0.467	1	0.533	1	0.469	1	0.168	1	1

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(2) 차원축소 배당률에 따른 랜덤 포레스트 승패예측모형 타당도

랜덤 포레스트 알고리즘을 기반으로 차원축소 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 배당률의 평균을 이용한 모형이 ACC(0.566), ERR(0.434), F-1(0.536), MCC(0.301)로 가장 높은 순위를 나타냈다. 두 번째는 과거 5경기 배당률의 평균을 이용한 모형이 ACC(0.555), ERR(0.445), F-1(0.513), MCC(0.273)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 28>과 같다.

표 28. 차원축소 배당률에 따른 랜덤 포레스트 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.544	3	0.456	3	0.487	5	0.250	4	4
2경기전	0.540	4	0.460	4	0.510	3	0.252	3	3
3경기전	0.566	1	0.434	1	0.536	1	0.301	1	1
4경기전	0.539	5	0.461	5	0.502	4	0.236	5	5
5경기전	0.555	2	0.445	2	0.513	2	0.273	2	2

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(3) 차원축소 배당률에 따른 XGBoost 승패예측모형 타당도

XGBoost 알고리즘을 기반으로 차원축소 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 1경기 배당률 자료를 이용한 모형이 ACC(0.563), ERR(0.437), F-1(0.495), MCC(0.286)로 가장 높은 순위를 나타냈다. 두 번째는 과거 4경기 배당률의 평균을 이용한 모형이 ACC(0.554), ERR(0.446), F-1(0.520), MCC(0.266)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 29>와 같다.

표 29. 차원축소 배당률에 따른 XGBoost 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.563	1	0.437	1	0.495	3	0.286	1	1
2경기전	0.535	4	0.465	4	0.479	5	0.235	4	4
3경기전	0.553	3	0.447	3	0.517	2	0.279	2	3
4경기전	0.554	2	0.446	2	0.520	1	0.266	3	2
5경기전	0.522	5	0.478	5	0.495	3	0.222	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(4) 차원축소 배당률에 따른 LightGBM 승패예측모형 타당도

LightGBM 알고리즘을 기반으로 차원축소 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 배당률의 평균을 이용한 모형이 ACC(0.550), ERR(0.450), F-1(0.493), MCC(0.270)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 배당률의 평균을 이용한 모형이 ACC(0.544), ERR(0.456), F-1(0.486), MCC(0.250)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 30>과 같다.

표 30. 차원축소 배당률에 따른 LightGBM 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.533	4	0.467	4	0.466	5	0.226	4	4
2경기전	0.544	2	0.456	2	0.486	2	0.250	2	2
3경기전	0.550	1	0.450	1	0.493	1	0.270	1	1
4경기전	0.527	5	0.473	5	0.472	3	0.202	5	5
5경기전	0.539	3	0.462	3	0.470	4	0.229	3	3

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(5) 배당률에 따른 로지스틱 회귀분석 승패예측모형 타당도

로지스틱 회귀분석 알고리즘을 기반으로 차원축소 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 4경기 배당률의 평균을 이용한 모형이 ACC(0.543), ERR(0.457), F-1(0.514), MCC(0.245)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 배당률의 평균을 이용한 모형이 ACC(0.537), ERR(0.463), F-1(0.497), MCC(0.241)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 31>과 같다.

표 31. 차원축소 배당률에 따른 로지스틱 회귀분석 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.525	3	0.475	3	0.486	3	0.221	3	3
2경기전	0.537	2	0.463	2	0.497	2	0.241	2	2
3경기전	0.512	4	0.488	5	0.478	5	0.209	4	4
4경기전	0.543	1	0.457	1	0.514	1	0.245	1	1
5경기전	0.506	5	0.475	4	0.486	4	0.197	5	4

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(6) 차원축소 배당률에 따른 서포트 벡터 머신 승패예측모형 타당도

서포트 벡터 머신 알고리즘을 기반으로 차원축소 과거 배당률 자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 4경기 배당률의 평균을 이용한 모형이 ACC(0.570), ERR(0.430), F-1(0.477), MCC(0.290)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 배당률의 평균을 이용한 모형이 ACC(0.548), ERR(0.452), F-1(0.458), MCC(0.260)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 배당률에 따른 타당도 검증결과는 다음 <표 32>와 같다.

표 32. 차원축소 배당률에 따른 서포트 벡터 머신 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.546	3	0.454	3	0.462	2	0.253	3	3
2경기전	0.548	2	0.452	2	0.458	3	0.260	2	2
3경기전	0.538	5	0.463	5	0.450	5	0.246	5	5
4경기전	0.570	1	0.430	1	0.477	1	0.290	1	1
5경기전	0.544	4	0.456	4	0.455	4	0.250	4	4

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(7) 차원축소 배당률 자료 기반 최적 승패예측모형 선택

차원축소 과정을 거친 배당률 자료 조합(1경기 전~5경기 전 평균)에 따른 6가지 분류 머신러닝 알고리즘의 승패예측모형 타당도를 검증하였다. 각 알고리즘의 최적 승패예측모형은 <표 33>에 나타냈으며, 최적 승패예측모형 간의 순위를 비교하였다. 그 결과 랜덤 포레스트 알고리즘을 이용한 승패예측모형(RM2)이 ACC(0.566), ERR(0.434), F-1(0.536), MCC(0.301)로 가장 높은 순위를 나타냈다. 두 번째는 서포트 벡터 머신 알고리즘을 이용한 승패예측모형(RM6)이 ACC(0.570), ERR(0.430), F-1(0.477), MCC(0.290)로 높은 순위를 기록하였다.

표 33. 차원축소 배당률에 따른 최적 승패예측모형 타당도 비교

모형	과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
RM1	5경기전	0.467	6	0.533	6	0.469	6	0.168	6	6
RM2	3경기전	0.566	2	0.434	2	0.536	1	0.301	1	1
RM3	1경기전	0.563	3	0.437	3	0.495	2	0.286	3	3
RM4	3경기전	0.550	4	0.450	4	0.493	3	0.270	4	4
RM5	1경기전	0.525	5	0.475	5	0.486	4	0.221	5	5
RM6	4경기전	0.570	1	0.430	1	0.477	5	0.290	2	2

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

· RM(Reduction Model)1: 차원축소 의사결정나무, · RM(Reduction Model)2: 차원축소 랜덤 포레스트,

· RM(Reduction Model)3: 차원축소 XGBoost, · RM(Reduction Model)4: 차원축소 LightGBM,

· RM(Reduction Model)5: 차원축소 로지스틱 회귀분석, · RM(Reduction Model)6: 차원축소 서포트 벡터 머신

3) 배당률 승패예측모형과 차원축소 배당률 승패예측모형 비교

배당률 승패예측모형과 차원축소 배당률 승패예측모형 간의 순위 비교 결과는 다음<표 34>와 같다. 차원축소 배당률 자료와 랜덤 포레스트 알고리즘을 함께 사용한 승패예측모형(RM2)이 ACC(0.566), ERR(0.434), F-1(0.536), MCC(0.301)로 가장 높은 순위로 나타났다. 1위부터 5위까지의 모형은 RM2> M5> M6> RM6> M3으로 나타났으며, 차원축소 자료를 사용한 승패예측모형과 기존 자료를 사용한 모형의 타당도 차이보다 분류 알고리즘 종류에 따른 타당도 차이가 크게 나타났다.

표 34. 최적 승패예측모형 종합 타당도 및 순위 비교

모형	과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
M1	3경기전	0.496	11	0.504	11	0.494	6	0.217	11	10
M2	3경기전	0.558	7	0.442	7	0.528	3	0.289	4	7
M3	1경기전	0.563	5	0.437	5	0.495	4	0.286	6	5
M4	3경기전	0.550	8	0.450	8	0.493	7	0.270	8	8
M5	4경기전	0.566	3	0.434	3	0.539	1	0.289	5	2
M6	4경기전	0.574	1	0.426	1	0.481	10	0.298	2	3
RM1	5경기전	0.467	12	0.533	12	0.469	12	0.168	12	12
RM2	3경기전	0.566	3	0.434	3	0.536	2	0.301	1	1
RM3	1경기전	0.563	5	0.437	5	0.495	4	0.286	6	5
RM4	3경기전	0.550	8	0.450	8	0.493	7	0.270	8	8
RM5	1경기전	0.525	10	0.475	10	0.486	9	0.221	10	10
RM6	4경기전	0.570	2	0.430	2	0.477	11	0.290	3	4

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

· M1: 의사결정나무, · M2: 랜덤 포레스트, · M3: XGBoost, · M4: LightGBM, · M5: 로지스틱 회귀분석, · M6: 서포트벡터머신

· RM(Reduction Model)1: 차원축소 의사결정나무, · RM(Reduction Model)2: 차원축소 랜덤 포레스트,

· RM(Reduction Model)3: 차원축소 XGBoost, · RM(Reduction Model)4: 차원축소 LightGBM,

· RM(Reduction Model)5: 차원축소 로지스틱 회귀분석, · RM(Reduction Model)6: 차원축소 서포트 벡터 머신

3. 혼합자료를 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교

1) 혼합자료를 활용한 승패예측모형 탐색

6가지 분류 머신러닝 알고리즘을 사용하여 잉글랜드 프리미어리그 승패예측모형을 탐색하였다. 사용된 입력 변인은 과거 경기자료와 배당률 자료를 통합한 혼합자료를 투입하였으며, 이전 1경기 자료부터 5경기 자료까지의 조합을 모두 분석하였다. 각 분류 알고리즘에 따라 어떤 조합의 혼합자료가 적합한지 알아보기 위하여 타당도 검증결과와 순위를 산출하였으며, 결과는 다음과 같다.

(1) 혼합자료에 따른 의사결정나무 승패예측모형 타당도

의사결정나무 알고리즘을 기반으로 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 5경기 혼합자료의 평균을 이용한 모형이 ACC(0.506), ERR(0.495), F-1(0.499), MCC(0.225)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 혼합자료의 평균을 이용한 모형이 ACC(0.495), ERR(0.506), F-1(0.494), MCC(0.218)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 35>와 같다.

표 35. 혼합자료에 따른 의사결정나무 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.476	3	0.524	3	0.472	3	0.193	3	3
2경기전	0.495	2	0.506	2	0.494	2	0.218	2	2
3경기전	0.465	4	0.535	4	0.468	4	0.177	4	4
4경기전	0.430	5	0.570	5	0.438	5	0.119	5	5
5경기전	0.506	1	0.495	1	0.499	1	0.225	1	1

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(2) 혼합자료에 따른 랜덤 포레스트 승패예측모형 타당도

랜덤 포레스트 알고리즘을 기반으로 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 4경기 혼합자료의 평균을 이용한 모형이 ACC(0.578), ERR(0.423), F-1(0.534), MCC(0.302)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기과 3경기 혼합자료의 평균을 이용한 모형으로 나타났다. 과거 2경기 혼합자료의 평균을 이용한 모형은 ACC(0.559), ERR(0.441), F-1(0.526), MCC(0.283)로 타당도가 검증되었으며, 과거 3경기 혼합자료의 평균을 이용한 모형은 ACC(0.558), ERR(0.442), F-1(0.528), MCC(0.289)로 확인되었다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 36>과 같다.

표 36. 혼합자료에 따른 랜덤 포레스트 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.555	4	0.445	4	0.503	4	0.273	4	4
2경기전	0.559	2	0.441	2	0.526	3	0.283	3	2
3경기전	0.558	3	0.442	3	0.528	2	0.289	2	2
4경기전	0.578	1	0.423	1	0.534	1	0.302	1	1
5경기전	0.539	5	0.462	5	0.500	5	0.251	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(3) 혼합자료에 따른 XGBoost 승패예측모형 타당도

XGBoost 알고리즘을 기반으로 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 1경기 혼합자료를 이용한 모형이 ACC(0.552), ERR(0.449), F-1(0.488), MCC(0.264)로 가장 높은 순위를 나타냈다. 두 번째는 과거 4경기 혼합자료의 평균을 이용한 모형이 ACC(0.543), ERR(0.457), F-1(0.496), MCC(0.236)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 37>과 같다.

표 37. 혼합자료에 따른 XGBoost 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.552	1	0.449	1	0.488	5	0.264	1	1
2경기전	0.537	3	0.463	3	0.492	3	0.242	3	4
3경기전	0.535	4	0.465	4	0.506	1	0.255	2	3
4경기전	0.543	2	0.457	2	0.496	2	0.236	4	2
5경기전	0.522	5	0.478	5	0.490	4	0.228	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(4) 혼합자료에 따른 LightGBM 승패예측모형 타당도

LightGBM 알고리즘을 기반으로 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 혼합자료의 평균을 이용한 모형이 ACC(0.563), ERR(0.437), F-1(0.494), MCC(0.295)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 혼합자료의 평균을 이용한 모형이 ACC(0.544), ERR(0.456), F-1(0.477), MCC(0.250)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 38>과 같다.

표 38. 혼합자료에 따른 LightGBM 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.535	5	0.465	5	0.479	2	0.232	4	4
2경기전	0.544	2	0.456	2	0.477	3	0.250	2	2
3경기전	0.563	1	0.437	1	0.494	1	0.295	1	1
4경기전	0.539	3	0.461	3	0.442	5	0.220	5	4
5경기전	0.538	4	0.462	4	0.454	4	0.232	3	3

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(5) 혼합자료에 따른 로지스틱 회귀분석 승패예측모형 타당도

로지스틱 회귀분석 알고리즘을 기반으로 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 4경기 혼합자료의 평균을 이용한 모형이 ACC(0.527), ERR(0.473), F-1(0.507), MCC(0.233)로 가장 높은 순위를 나타냈다. 두 번째는 과거 1경기 혼합자료를 이용한 모형이 ACC(0.519), ERR(0.481), F-1(0.489), MCC(0.216)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 39>와 같다.

표 39. 혼합자료에 따른 로지스틱 회귀분석 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.519	2	0.481	2	0.489	3	0.216	2	2
2경기전	0.515	3	0.485	3	0.490	2	0.211	4	3
3경기전	0.509	4	0.491	4	0.486	4	0.212	3	4
4경기전	0.527	1	0.473	1	0.507	1	0.233	1	1
5경기전	0.451	5	0.550	5	0.443	5	0.121	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(6) 혼합자료에 따른 서포트 벡터 머신 승패예측모형 타당도

서포트 벡터 머신 알고리즘을 기반으로 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 5경기 혼합자료를 이용한 모형이 ACC(0.546), ERR(0.454), F-1(0.464), MCC(0.252)로 가장 높은 순위를 나타냈다. 두 번째는 과거 3경기 혼합자료의 평균을 이용한 모형이 ACC(0.550), ERR(0.450), F-1(0.465), MCC(0.272)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 40>과 같다.

표 40. 혼합자료에 따른 서포트 벡터 머신 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.546	3	0.454	3	0.464	3	0.252	3	3
2경기전	0.542	4	0.458	4	0.456	4	0.248	4	4
3경기전	0.550	2	0.450	2	0.465	2	0.272	1	2
4경기전	0.554	1	0.446	1	0.466	1	0.255	2	1
5경기전	0.528	5	0.473	5	0.449	5	0.222	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(7) 혼합자료 기반 최적 승패예측모형 선택

혼합자료 조합(1경기 전~5경기 전 평균)에 따른 6가지 분류 머신러닝 알고리즘의 승패 예측모형 타당도를 검증하였다. 각 알고리즘의 최적 승패예측모형은 <표 41>에 나타냈으며, 최적 승패예측모형 간의 순위를 비교하였다. 그 결과 랜덤 포레스트 알고리즘을 이용한 승패예측모형(M2)이 ACC(0.578), ERR(0.423), F-1(0.534), MCC(0.302)로 가장 높은 순위를 나타냈다. 두 번째는 LightGBM 알고리즘을 이용한 승패예측모형(M4)이 ACC(0.563), ERR(0.437), F-1(0.494), MCC(0.295)로 높은 순위를 기록하였다.

표 41. 혼합자료에 따른 최적 승패예측모형 타당도 비교

모형	과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
M1	5경기전	0.506	6	0.495	6	0.499	3	0.225	6	6
M2	4경기전	0.578	1	0.423	1	0.534	1	0.302	1	1
M3	1경기전	0.552	4	0.449	4	0.488	5	0.264	3	3
M4	3경기전	0.563	2	0.437	2	0.494	4	0.295	2	2
M5	4경기전	0.527	5	0.473	5	0.507	2	0.233	5	5
M6	4경기전	0.554	3	0.446	3	0.466	6	0.255	4	3

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

· M1: 의사결정나무, · M2: 랜덤 포레스트, · M3: XGBoost, · M4: LightGBM, · M5: 로지스틱 회귀분석, · M6: 서포트벡터머신

2) 차원축소 혼합자료를 활용한 승패예측모형 탐색

연구내용 3-2에서는 차원축소 과정을 통해 변인이 축소된 혼합자료를 활용한 승패예측모형을 탐색하였다. 변인 중요도에 따른 선택된 변인은 상위 12개로 한정하였으며, 자료의 조합에 따른 변인 항목은 부록에 첨부하였다. 각 분류 머신러닝 알고리즘의 타당도 검증결과와 순위는 다음과 같다.

(1) 차원축소 혼합자료에 따른 의사결정나무 승패예측모형 타당도

의사결정나무 알고리즘을 기반으로 차원축소 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 혼합자료의 평균을 이용한 모형이 ACC(0.522), ERR(0.478), F-1(0.523), MCC(0.260)로 가장 높은 순위를 나타냈다. 두 번째는 과거 5경기 혼합자료의 평균을 이용한 모형이 ACC(0.478), ERR(0.522), F-1(0.467), MCC(0.165)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 42>와 같다.

표 42. 차원축소 혼합자료에 따른 의사결정나무 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.446	4	0.554	4	0.443	5	0.138	4	4
2경기전	0.463	3	0.537	3	0.457	3	0.152	3	3
3경기전	0.522	1	0.478	1	0.523	1	0.260	1	1
4경기전	0.446	5	0.554	5	0.447	4	0.129	5	5
5경기전	0.478	2	0.522	2	0.467	2	0.165	2	2

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(2) 차원축소 혼합자료에 따른 랜덤 포레스트 승패예측모형 타당도

랜덤 포레스트 알고리즘을 기반으로 차원축소 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 혼합자료의 평균을 이용한 모형이 ACC(0.558), ERR(0.442), F-1(0.528), MCC(0.289)로 가장 높은 순위를 나타냈다. 두 번째는 과거 5경기 혼합자료의 평균을 이용한 모형이 ACC(0.550), ERR(0.450), F-1(0.511), MCC(0.266)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 43>과 같다.

표 43. 차원축소 혼합자료에 따른 랜덤 포레스트 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.536	4	0.464	4	0.482	5	0.237	4	4
2경기전	0.542	3	0.458	3	0.511	2	0.254	3	3
3경기전	0.558	1	0.442	1	0.528	1	0.289	1	1
4경기전	0.535	5	0.465	5	0.499	4	0.230	5	5
5경기전	0.550	2	0.450	2	0.511	3	0.266	2	2

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(3) 차원축소 혼합자료에 따른 XGBoost 승패예측모형 타당도

XGBoost 알고리즘을 기반으로 차원축소 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 1경기 혼합자료를 이용한 모형이 ACC(0.552), ERR(0.449), F-1(0.488), MCC(0.264)로 가장 높은 순위를 나타냈다. 두 번째는 과거 4경기 혼합자료의 평균을 이용한 모형이 ACC(0.543), ERR(0.457), F-1(0.496), MCC(0.236)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 44>와 같다.

표 44. 차원축소 혼합자료에 따른 XGBoost 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.552	1	0.449	1	0.488	5	0.264	1	1
2경기전	0.537	3	0.463	3	0.492	3	0.242	3	4
3경기전	0.535	4	0.465	4	0.507	1	0.255	2	3
4경기전	0.543	2	0.457	2	0.496	2	0.236	4	2
5경기전	0.522	5	0.478	5	0.490	4	0.228	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(4) 차원축소 혼합자료에 따른 LightGBM 승패예측모형 타당도

LightGBM 알고리즘을 기반으로 차원축소 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 3경기 혼합자료의 평균을 이용한 모형이 ACC(0.563), ERR(0.437), F-1(0.494), MCC(0.295)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 혼합자료의 평균을 이용한 모형이 ACC(0.544), ERR(0.456), F-1(0.477), MCC(0.250)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 45>와 같다.

표 45. 차원축소 혼합자료에 따른 LightGBM 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.535	5	0.465	5	0.479	2	0.232	4	4
2경기전	0.544	2	0.456	2	0.477	3	0.250	2	2
3경기전	0.563	1	0.437	1	0.494	1	0.295	1	1
4경기전	0.539	3	0.461	3	0.442	5	0.220	5	4
5경기전	0.539	4	0.462	4	0.454	4	0.232	3	3

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(5) 차원축소 혼합자료에 따른 로지스틱 회귀분석 승패예측모형 타당도

로지스틱 회귀분석 알고리즘을 기반으로 차원축소 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 4경기 혼합자료의 평균을 이용한 모형이 ACC(0.543), ERR(0.457), F-1(0.514), MCC(0.245)로 가장 높은 순위를 나타냈다. 두 번째는 과거 2경기 혼합자료의 평균을 이용한 모형이 ACC(0.535), ERR(0.465), F-1(0.496), MCC(0.238)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 46>과 같다.

표 46. 차원축소 혼합자료에 따른 로지스틱 회귀분석 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.523	3	0.477	4	0.482	4	0.215	4	3
2경기전	0.535	2	0.465	2	0.496	2	0.238	2	2
3경기전	0.519	4	0.481	5	0.488	3	0.224	3	3
4경기전	0.543	1	0.457	1	0.514	1	0.245	1	1
5경기전	0.500	5	0.475	3	0.452	5	0.187	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(6) 차원축소 혼합자료에 따른 서포트 벡터 머신 승패예측모형 타당도

서포트 벡터 머신 알고리즘을 기반으로 차원축소 과거 혼합자료의 조합들을 적용하여 타당도를 검증하고 모형별 순위를 산출하였다. 그 결과 과거 4경기 혼합자료의 평균을 이용한 모형이 ACC(0.574), ERR(0.426), F-1(0.481), MCC(0.298)로 가장 높은 순위를 나타냈다. 두 번째는 과거 3경기 혼합자료의 평균을 이용한 모형이 ACC(0.548), ERR(0.452), F-1(0.464), MCC(0.267)로 높은 순위를 나타냈다. 구체적인 과거 경기(1경기 전, 2경기 전, 3경기 전, 4경기 전, 5경기 전) 혼합자료에 따른 타당도 검증결과는 다음 <표 47>과 같다.

표 47. 차원축소 혼합자료에 따른 서포트 벡터 머신 승패예측모형 타당도

과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
1경기전	0.545	4	0.455	4	0.461	4	0.250	4	4
2경기전	0.546	3	0.454	3	0.456	5	0.256	3	3
3경기전	0.548	2	0.452	2	0.464	3	0.267	2	2
4경기전	0.574	1	0.426	1	0.481	1	0.298	1	1
5경기전	0.544	5	0.456	5	0.473	2	0.249	5	5

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

(7) 차원축소 혼합자료 기반 최적 승패예측모형 선택

차원축소 과정을 거친 혼합자료 조합(1경기 전~5경기 전 평균)에 따른 6가지 분류 머신 러닝 알고리즘의 승패예측모형 타당도를 검증하였다. 각 알고리즘의 최적 승패예측모형은 <표 48>에 나타냈으며, 최적 승패예측모형 간의 순위를 비교하였다. 그 결과 서포트 벡터 머신 알고리즘을 이용한 승패예측모형(RM6)이 ACC(0.574), ERR(0.426), F-1(0.481), MCC(0.298)로 가장 높은 순위를 나타냈다. 두 번째는 LightGBM 알고리즘을 이용한 승패예측모형(RM4)이 ACC(0.563), ERR(0.437), F-1(0.494), MCC(0.295)로 높은 순위를 기록하였다.

표 48. 차원축소 혼합자료에 따른 최적 승패예측모형 타당도 비교

모형	과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
RM1	3경기전	0.522	6	0.478	6	0.523	2	0.260	5	5
RM2	3경기전	0.558	3	0.442	3	0.528	1	0.289	3	2
RM3	1경기전	0.552	4	0.449	4	0.488	5	0.264	4	4
RM4	3경기전	0.563	2	0.437	2	0.494	4	0.295	2	2
RM5	4경기전	0.543	5	0.457	5	0.514	3	0.245	6	5
RM6	4경기전	0.574	1	0.426	1	0.481	6	0.298	1	1

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

· RM(Reduction Model)1: 차원축소 의사결정나무, · RM(Reduction Model)2: 차원축소 랜덤 포레스트,

· RM(Reduction Model)3: 차원축소 XGBoost, · RM(Reduction Model)4: 차원축소 LightGBM,

· RM(Reduction Model)5: 차원축소 로지스틱 회귀분석, · RM(Reduction Model)6: 차원축소 서포트 벡터 머신

3) 혼합자료 승패예측모형과 차원축소 혼합자료 승패예측모형 비교

혼합자료 승패예측모형과 차원축소 혼합자료 승패예측모형 간의 순위 비교 결과는 다음 <표 49>와 같다. 혼합자료와 랜덤 포레스트 알고리즘을 함께 사용한 승패예측모형(M2)이 ACC(0.578), ERR(0.423), F-1(0.534), MCC(0.302)로 가장 높은 순위로 나타났다. 1위부터 5위까지의 모형은 M2> RM4, M4> RM2, RM6으로 나타났으며, 차원축소 자료를 사용한 승패예측모형과 기존 자료를 사용한 모형의 타당도 차이보다 분류 알고리즘 종류에 따른 타당도 차이가 크게 나타났다.

표 49. 최적 승패예측모형 종합 타당도 및 순위 비교

모형	과거 N 경기	ACC	순위	ERR	순위	F-1	순위	MCC	순위	전체 순위
M1	5경기전	0.506	12	0.495	12	0.499	6	0.225	12	12
M2	4경기전	0.578	1	0.423	1	0.534	1	0.302	1	1
M3	1경기전	0.552	7	0.449	7	0.488	9	0.264	6	6
M4	3경기전	0.563	3	0.437	3	0.494	7	0.295	3	2
M5	4경기전	0.527	10	0.473	10	0.507	5	0.233	11	11
M6	4경기전	0.554	6	0.446	6	0.466	12	0.255	9	9
RM1	3경기전	0.522	11	0.478	11	0.523	3	0.260	8	9
RM2	3경기전	0.558	5	0.442	5	0.528	2	0.289	5	4
RM3	1경기전	0.552	7	0.449	7	0.488	9	0.264	6	6
RM4	3경기전	0.563	3	0.437	3	0.494	7	0.295	3	2
RM5	4경기전	0.543	9	0.457	9	0.514	4	0.245	10	8
RM6	4경기전	0.574	2	0.426	2	0.481	11	0.298	2	4

· ACC: Accuracy, ERR: Error rate, F-1: F-1 score, MCC: Matthews correlation coefficient

· M1: 의사결정나무, · M2: 랜덤 포레스트, · M3: XGBoost, · M4: LightGBM, · M5: 로지스틱 회귀분석, · M6: 서포트벡터머신

· RM(Reduction Model)1: 차원축소 의사결정나무, · RM(Reduction Model)2: 차원축소 랜덤 포레스트,

· RM(Reduction Model)3: 차원축소 XGBoost, · RM(Reduction Model)4: 차원축소 LightGBM,

· RM(Reduction Model)5: 차원축소 로지스틱 회귀분석, · RM(Reduction Model)6: 차원축소 서포트 벡터 머신

V. 논의

축구 경기결과를 예측하려는 노력(김주학 등, 2007; 최형준, 이윤수, 2019; 홍종선, 정민섭, 이재형, 2010; Berrar, Lopes, & Dubitzky, 2019)은 지속되고 있으며, 다양한 머신러닝 알고리즘과 자료를 통한 축구 승패예측모형 구축은 발전되어왔다. 그러나 자료의 특성에 맞는 효과적인 알고리즘을 결정하는데 많은 시행착오를 겪는다는 문제와(이영섭, 오현정, 김미경, 2005) 불필요한 가정에 의한 모형의 복잡성 문제(Domingos, 1999)가 제기되어 왔다.

구체적으로 선행연구에서 축구 승패예측모형에 사용된 머신러닝 알고리즘과 자료의 형태를 살펴보면 다음과 같다. 승패예측모형에 사용된 머신러닝 알고리즘은 주로 분류 알고리즘(의사결정나무, 랜덤 포레스트, XGBoost, LightGMB, 로지스틱 회귀분석, 서포트 벡터 머신)이 사용되었으며, 분석에 사용된 자료의 형태는 경기기록, 배당률, 경기기록과 배당률을 통합한 혼합자료로 구분되어 있다. 그러나 분류 알고리즘과 각각의 자료 형태를 동시에 비교 분석한 연구가 미흡하기 때문에 다양한 조합의 모형을 탐색할 필요성이 있다고 판단하였다. 또한, 차원이 높은 복잡한 자료를 사용함에 따라 해석과 일반화가 어렵다는 문제점이 강조되고 있어 자료의 차원을 축소하는 방법에 대한 필요성이 언급되어 왔다. 이에 이 연구에서는 원자료와 차원축소 자료에 따른 각 모형을 비교하였다.

따라서 이 연구의 목적은 잉글랜드 프리미어리그 경기기록과 배당률 자료를 활용하여 분류 머신러닝 알고리즘 기반 축구 승패예측모형을 탐색 및 비교하는 것이다. 연구의 목적을 달성하기 위한 연구내용은 다음과 같다. 첫째, 원자료 경기기록과 차원축소 경기기록을 활용한 승패예측모형들을 탐색 및 비교하였다. 둘째, 원자료 배당률과 차원축소 배당률을 활용한 승패예측모형들을 탐색 및 비교하였다. 셋째, 원자료 혼합자료와 차원축소 혼합자료를 활용한 승패예측모형들을 탐색 및 비교하였다. 각 연구내용의 세부내용은 1) 원자료를 활용한 6가지 분류 머신러닝 알고리즘 기반 승패예측모형 탐색, 2) 차원축소 자료를 활용한 6가지 분류 머신러닝 알고리즘 기반 승패예측모형 탐색, 3) 원자료 승패예측모형과 차원축소 승패예측모형 간의 비교를 하였다. 이러한 과정을 통해 분류 머신러닝 알고리즘

과 자료 형태에 따라 승패예측모형의 타당도가 어떻게 변화하는지 확인하였다. 이 연구를 통해 얻은 연구결과별 논의는 다음과 같다.

1. 경기기록을 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교

첫 번째 연구내용에서는 경기기록과 분류 머신러닝 알고리즘을 활용하여 축구 승패예측모형을 탐색하고 비교하였다. 과거 경기기록을 입력 변인으로 투입하였으며, 이전 1경기 자료부터 5경기 자료까지 모든 조합을 살펴보았다. 6가지 분류 머신러닝 알고리즘과 자료 조합들을 분석한 결과 최적 승패예측모형의 조합은 다음과 같다. 차원축소 경기기록, 랜덤 포레스트와 과거 3경기 자료의 평균을 이용한 모형(RM2)이 ACC(0.548), ERR(0.452), F-1(0.470), MCC(0.270)로 가장 높은 순위를 기록하였다. RM2 모형의 경기결과 예측 삼원분류표는 다음 <그림 9>와 같다.

		실제 경기결과		
		어웨이	무승부	홈
승패예측모형	어웨이	65 (16.8%)	0 (0.0%)	49 (12.7%)
	무승부	29 (7.5%)	1 (0.3%)	66 (17.1%)
	홈	31 (8.0%)	0 (0.0%)	146 (37.7%)

그림 9. 경기결과 예측 삼원분류표

분류표에서 볼 수 있듯이 전반적으로 무승부 결과에 대한 예측력이 낮다는 것을 확인할 수 있다. 김형원(2020)은 축구 승패예측모형을 개발하는 것에서 균형 잡힌 성능 분포를 나타내는 것은 어려운 일이라고 강조하면서 무승부 경기를 예측하는데 더 다양한 변인과 모형구조가 제안되어야 한다고 강조했다. 후속연구에서는 홈승, 어웨이승 뿐만 아니라 무승

부 경기를 정확하게 예측할 수 있는 연구가 이루어지길 기대한다.

2. 배당률을 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교

두 번째 연구내용에서는 배당률과 분류 머신러닝 알고리즘을 활용하여 축구 승패예측모형을 탐색하고 비교하였다. 배당률을 입력 변인으로 투입하였으며, 이전 1경기 자료부터 5경기 자료까지 모든 조합을 살펴보았다. 6가지 분류 머신러닝 알고리즘과 자료 조합들을 분석한 결과 최적 승패예측모형의 조합은 다음과 같다. 차원축소 배당률 자료, 랜덤 포레스트와 과거 3경기 자료의 평균을 이용한 모형(RM2)이 ACC(0.566), ERR(0.434), F-1(0.536), MCC(0.301)로 가장 높은 순위를 기록하였다. 배팅 회사들의 배당률 자료는 서로 높은 관련성이 있어 다중공선성(multicollinearity)의 문제를 유발할 수 있으며 이는 낮은 예측 정확도의 원인이 될 수 있다(Tak, & Joustra, 2015). 하지만 머신러닝에서는 독립변인 간의 상관성이 높더라도 랜덤 포레스트 알고리즘과 같이 정규화하는 기능이 있어 문제를 해결할 수 있다(Deng, & Runger, 2013). 이러한 사실은 배당률에 따른 최적 예측모형이 랜덤 포레스트 알고리즘이라는 결과를 지지해주고 있다.

3. 혼합자료를 활용한 분류 머신러닝 알고리즘 기반 승패예측모형 탐색 및 비교

세 번째 연구내용에서는 혼합자료와 분류 머신러닝 알고리즘을 활용하여 축구 승패예측모형을 탐색하고 비교하였다. 혼합자료를 입력 변인으로 투입하였으며, 이전 1경기 자료부터 5경기 자료까지 모든 조합을 살펴보았다. 6가지 분류 머신러닝 알고리즘과 자료 조합들을 분석한 결과 최적 승패예측모형의 조합은 다음과 같다. 혼합자료, 랜덤 포레스트와 과거 4경기 자료의 평균을 이용한 모형(M2)이 ACC(0.578), ERR(0.423), F-1(0.534), MCC(0.302)로 가장 높은 순위를 기록하였다. 혼합자료를 사용하는 것이 다른 자료 형태보다 예측 성능이 높은 것을 확인할 수 있었으며, 무승부에 대해서도 예측 정확도가 증가하였다. 예측

정확도는 약 58%로 기존의 연구(Hoekstra, Bison, & Eiben, 2012; Ulmer, Fernandez, & Peterson, 2013)와 비슷한 수준임을 확인할 수 있었다. 하지만 기존의 변인을 조합하여 새로운 변인을 만드는 피처 엔지니어링(feature engineering) 과정을 거친 연구(Baboota & Kaur, 2019; Shin & Gasparyan, 2014)에서는 예측 정확도가 증가하였다. 후속연구에서는 피처 엔지니어링 방법을 통해 모형의 정확도를 상승시키는 새로운 경기기록 변인에 관한 연구가 필요하다고 판단된다.

종합적으로, 이 연구에서는 다양한 자료 형태에 따른 축구 승패예측모형을 탐색하고 비교하였다. 하지만 기존 변인에 대한 피처 엔지니어링 방법을 적용하지 못했다는 점과 무승부 예측에 한계가 있었다. 그럼에도 이 연구를 통해 밝혀낸 축구 승패예측모형 탐색 및 비교 결과는 다양한 분야에서 실제로 승패예측모형을 구축할 때 기초자료로써 활용할 수 있을 것이다.

Ⅵ. 결론

이 연구의 목적은 잉글랜드 프리미어리그 경기기록과 배당률 자료를 활용하여 분류 머신러닝 알고리즘 기반 축구 승패예측모형을 탐색 및 비교하는 것이다. 이 연구의 내용은 크게 세 부분으로 나누어진다. 첫째, 원자료 경기기록과 차원축소 경기기록을 활용한 승패예측모형들을 탐색 및 비교한다. 둘째, 원자료 배당률과 차원축소 배당률을 활용한 승패예측모형들을 탐색 및 비교한다. 셋째, 원자료 혼합자료와 차원축소 혼합자료를 활용한 승패예측모형들을 탐색 및 비교하는 것이다. 각 연구내용의 세부내용은 1) 원자료를 활용한 6가지 분류 머신러닝 알고리즘 기반 승패예측모형 탐색, 2) 차원축소 자료를 활용한 6가지 분류 머신러닝 알고리즘 기반 승패예측모형 탐색, 3) 원자료 승패예측모형과 차원축소 승패예측모형 간의 비교였다. 이러한 과정을 통해 분류 머신러닝 알고리즘과 자료의 형태에 따라 승패예측모형의 타당도가 어떻게 변화하는지 확인하였다. 이 연구의 결론은 다음과 같다.

첫째, 원자료 경기기록과 차원축소 경기기록을 활용한 승패예측모형들을 탐색하고 비교한 결과 차원축소 경기기록 자료와 랜덤 포레스트 알고리즘을 함께 사용한 승패예측모형(RM2)이 분류정확도(Accuracy) 54.8%로 가장 높은 순위로 나타났다.

둘째, 원자료 배당률과 차원축소 배당률을 활용한 승패예측모형들을 탐색하고 비교한 결과 차원축소 배당률 자료와 랜덤 포레스트 알고리즘을 함께 사용한 승패예측모형(RM2)이 분류정확도 56.6%로 가장 높은 순위로 나타났다.

셋째, 원자료 혼합자료와 차원축소 혼합자료를 활용한 승패예측모형들을 탐색하고 비교한 결과 혼합자료와 랜덤 포레스트 알고리즘을 함께 사용한 승패예측모형(M2)이 분류정확도 57.8%로 가장 높은 순위로 나타났다.

결론적으로 이 연구에서는 잉글랜드 프리미어리그 경기기록과 배당률 자료를 활용하여 분류 머신러닝 알고리즘 기반 축구 승패예측모형을 탐색 및 비교하였다. 이는 머신러닝을 활용하여 축구 승패예측모형을 구축할 때 자료 형태에 따른 분류 머신러닝 알고리즘 선택의 기초정보로 활용될 수 있을 것으로 기대한다.

참고문헌

- 강상조 (1994). **체육통계**. 서울: 도서출판, 21.
- 김세형, 강상조, 박재현, 김혜진 (2008). 한국프로농구 경기기록 분석에 의한 승패결정요인. **한국체육측정평가학회지**, 10(1), 1-12.
- 김인호, 이경섭 (2020). 트리 기반 앙상블 방법을 활용한 자동 평가 모형 개발 및 평가: 서울특별시 주거용 아파트를 사례로. **한국데이터정보과학회지**, 31(2), 375-389.
- 김주학, 노갑택, 박종성, 이원희 (2007). 신경망분석을 이용한 축구경기 승, 패 예측모형 개발-2006 독일월드컵 대회를 중심으로. **체육과학연구**, 18(4), 54-63.
- 김혁 (2019). 기계학습 방법에 의한 프로스포츠에서의 관중 수 예측과 그 요인들 연구. **Journal of The Korean Data Analysis Society**, 21(4), 1867-1880.
- 김형원 (2020). 익스트림 그래디언트 부스팅 알고리즘에 기반한 축구 경기 예측. 서강대학교 정보통신대학원.
- 박재현, 강민수, 이진오, 강상조 (2005). 체육측정평가학: 결측치 처리: 어떤 방법이 최선인가?. **한국체육학회지**, 44(1), 385-398.
- 박홍진 (2020). ‘인공지능’, ‘기계학습’, ‘딥 러닝’ 분야의 국내 논문 동향 분석. **한국정보전자통신기술학회 논문지**, 13(4), 283-292.
- 송상운 (2015). 예대금리차 결정요인 모형의 예측력 비교 연구-Ridge, LASSO 및 Elastic Net 방법론을 중심으로. **금융지식연구**, 13(3), 41-65.
- 유진은 (2015). 랜덤 포레스트. **교육평가연구**, 28, 427-448.
- 윤보람 (2020). 인사 자료 분석에서 이직 분류를 위한 기계 학습의 활용. 서울대학교 대학원.
- 이석원, 천영진 (2017). 다중회귀분석을 이용한 메이저리그 승률의 모형구축과 예측. **Journal of The Korean Data Analysis Society**, 19(4), 2071-2079.
- 이영섭, 오현정, 김미경 (2005). 데이터 마이닝에서배경, 부스팅, SVM 분류 알고리즘 비

- 교 분석. **응용통계연구**, 18(2), 343-354.
- 이장택 (2020). 분위수 회귀모형을 이용한 한국프로야구 투수들의 연봉 결정요인. **한국데이터정보과학회지**, 31(4), 653-662.
- 이재현, 이수원 (2020). 앙상블 기법을 통한 잉글리시 프리미어리그 경기결과 예측. **정보처리학회논문지. 소프트웨어 및 데이터 공학**, 9(5), 161-168.
- 정인범 (2018). 통계적 해석 방법과 모델 기반 방법을 사용한 차원축소: Elementary effect 기법과 random forest regressor 의 비교. 한양대학교 대학원.
- 조준모 (2019). 빅데이터의 정규화 전처리과정이 기계학습의 성능에 미치는 영향. **한국전자통신학회 논문지**, 14, 547-552.
- 채진석, 조은형, 엄한주 (2010). 프로야구 포스트시즌 진출 예측을 위한 통계적 모형 비교. **한국체육측정평가학회지**, 12(1), 33-48.
- 최재일, 정용락 (2010). 시계열 분석에 의한 한국 프로축구 관중수 예측 (2009-2015). **한국사회체육학회지**, 39(2), 921-928.
- 최창환, 윤지운 (2017). 경륜 출주정보를 활용한 승자 예측모형 탐색: 데이터마이닝 기반 의사결정나무분석의 적용. **한국체육측정평가학회지**, 19(4), 15-26.
- 최형준, 이운수 (2019). 축구 월드컵대회의 경기기록 기반 경기결과 예측. **한국체육과학회지**, 28(1), 1317-1325.
- 하대우, 김영민, 안재준 (2019). XGBoost 모형을 활용한 코스피 200 주가지수 등락 예측에 관한 연구. **한국데이터정보과학회지**, 30(3), 655-669.
- 한필수, 이석인 (2003). 스포츠산업, 경영: 2001-2002 시즌 데이터를 이용한 프로농구선수의 연봉모형 예측 및 변수생성. **한국체육학회지**, 42(3), 477-486.
- 홍종선, 정민섭, 이재형 (2010). 2010 남아공 월드컵 축구 예측모형 분석. **한국데이터정보과학회지**, 21(6), 1137-1146.
- Arabzad, S. M., Tayebi Araghi, M. E., Sadi-Nezhad, S., & Ghofrani, N. (2014). Football match results prediction using artificial neural networks; the case of Iran Pro

- League. *Journal of Applied Research on Industrial Engineering*, 13), 159-179.
- Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, 35(2), 741-755.
- Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine learning*, 108(1), 97-126.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.
- Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's razor. *Information processing letters*, 24(6), 377-380.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.
- Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, 44(1), 108-132.
- Buursma, D. (2011). Predicting sports events from past results Towards effective betting on football matches. *In Conference Paper, presented at 14th Twente Student Conference on IT*, Twente, Holland (Vol. 21).
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- Chui, M. (2017). Artificial intelligence the next digital frontier?. *McKinsey and Company Global Institute*, 47, 3-6.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

- Deng, H., & Runger, G. (2013). Gene selection with guided regularized random forest. *Pattern Recognition*, 46(12), 3483–3489.
- Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2), 265–280.
- Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data mining and knowledge discovery*, 3(4), 409–425.
- Domingues, J., Lopes, B., Mihaylova, P., & Georgieva, P. (2019). Incremental Learning for Football Match Outcomes Prediction. *In Iberian Conference on Pattern Recognition and Image Analysis* (pp. 217–228). Springer, Cham.
- Gevaria, K., Sanghavi, H., Vadiya, S., & Deulkar, K. (2015). Football match winner prediction. *International Journal of Emerging Technology and Advanced Engineering*, 5(10).
- Hoekstra, V., Bison, P., & Eiben, G. (2012). *Predicting football results with an evolutionary ensemble classifier*. Master's thesis, VU University Amsterdam, Amsterdam, The Netherlands.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression (Vol. 398)*. John Wiley & Sons.
- Hubáček, O., Šourek, G., & Železný, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, 108(1), 29–47.
- Igiri, C. P., & Nwachukwu, E. O. (2014). An improved prediction system for football a match result. *IOSR Journal of Engineering (IOSRJEN)*, 4(12), 12–20.
- Iskandaryan, D., Ramos, F., Palinggi, D. A., & Trilles, S. (2020). The Effect of Weather in Soccer Results: An Approach Using Machine Learning Techniques. *Applied Sciences*, 10(19), 6750.

- Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems, 19*(7), 544-553.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *In Advances in neural information processing systems* (pp. 3146-3154).
- Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 16*(1), 3-24.
- Lin, C., Wu, D., Liu, H., Xia, X., & Bhattarai, N. (2020). Factor Identification and Prediction for Teen Driver Crash Severity Using Machine Learning: A Case Study. *Applied Sciences, 10*(5), 1675.
- Linstone, H. A. (2002). Corporate planning, forecasting, and the long wave. *Futures, 34*(3-4), 317-336.
- Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica, 36*(3), 109-118.
- Min, B., Kim, J., Choe, C., Eom, H., & McKay, R. B. (2008). A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems, 21*(7), 551-562.
- Muhammad, I., & Yan, Z. (2015). SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *ICTACT Journal on Soft Computing, 5*(3).
- Nivard, W., & Mei, R. D. (2012). *Soccer analytics: Predicting the of soccer matches*. University of Amsterdam.
- Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). Football result prediction with Bayesian network in Spanish League-Barcelona team. *International Journal of*

- Computer Theory and Engineering*, 55), 812.
- Prasetio, D. (2016, August). Predicting football match results with logistic regression. *In 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)* (pp. 1-5). IEEE.
- Raheel, S. (2018). *Choosing the right encoding method-Label vs One hot encoder*: Towards datascience.
- RColorBrewer, S., & Liaw, M. A. (2018). *Package 'randomForest'*. University of California, Berkeley: Berkeley, CA, USA.
- Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3), 399-418.
- Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4), 13-22.
- Shin, J., & Gasparyan, R. (2014). *A novel way to soccer match prediction*. Stanford University: Department of Computer Science.
- Spann, M., & Skiera, B. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 28(1), 55-72.
- Štrumbelj, E., & Šikonja, M. R. (2010). Online bookmakers' odds as forecasts: The case of European soccer leagues. *International Journal of Forecasting*, 26(3), 482-488.
- Stübinger, J., Mangold, B., & Knoll, J. (2020). Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics. *Applied Sciences*, 10(1), 46.
- Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A

- review. *Data classification: Algorithms and applications*, 37.
- Tax, N., & Joustra, Y. (2015). Predicting the Dutch football competition using public data: A machine learning approach. *Trans. Knowl. Data Eng.*, 10 (10), 1-13.
- Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.
- Tiwari, T., Tiwari, T., & Tiwari, S. (2018). How Artificial Intelligence, Machine Learning and Deep Learning are Radically Different?. *International journal of Advanced Research in Computer Science and Software Engineering*, 8(2), 1-9.
- Ulmer, B., Fernandez, M., & Peterson, M. (2013). *Predicting soccer match results in the English Premier League* (Doctoral dissertation, Ph. D. thesis, Doctoral dissertation, Ph. D. dissertation, Stanford).
- Wunderlich, F., & Memmert, D. (2018). The Betting Odds Rating System: Using soccer forecasts to forecast soccer. *PloS one*, 13(6), e0198668.
- Yom-Tov, E. (2003). An introduction to pattern classification. *In Summer School on Machine Learning* (pp. 1-20). Springer, Berlin, Heidelberg.
- Zhang, C., & Ma, Y. (Eds.). (2012). Ensemble machine learning: methods and applications. *Springer Science & Business Media*.

ABSTRACT

Exploring England Premier League Win / Loss Prediction Model using Classification Machine Learning

Seung-Bak Lee

(Graduate School of Korea National Sport University)

Jae-Hyeon Park(Advisor)

The purpose of this study is to use English Premier League game records and odds to explore and compare soccer Win/Loss prediction models based on classification machine learning algorithm. The research is divided into three parts of achieve its purpose. The first research is 'Discovery and Comparison of Classification Machine Learning-Based Win/Loss Prediction Models Using Competition Records.' For this purpose, 44 variables were selected for analysis from the game records of 3,800 games from the 09-10 to 18-19 seasons of the English Premier League which were collected from English Premier League official website (www.premierleague.com) and whoscored.com (whoscored.com). The previous competition average was used as the input variable to predict the results after testing and comparing for all previous 1-5 competitions to identify an effective combination in the analysis. The second research is

'Discovery and Comparison of Win/Loss Prediction Models Based on Classification Machine Learning Algorithms Using Odds.' For this purpose, 36 variables were selected from the odds for the games collected from football-data. The third research is 'Discovery and Comparison of Win/Loss Prediction Models Based on Classification Machine Learning Algorithms Using Mixed Data.' To this end, all the previously collected competition records and odds data were integrated, and analyzed 83 variables, including common variables that inform the results of the competition and the team name, to examine the optimal Win/Loss prediction model. The conclusions of this study are as follows.

First, as a result of exploring and comparing the Win/Loss prediction models using the raw material competition records and the dimensional reduction competition records, the Win/Loss prediction model(RM2) using both the dimensional reduction competition record data and the random forest algorithm were ranked highest with accuracy 54.8%.

Second, as a result of exploring and comparing Win/Loss prediction models using raw material odds and dimension reduction odds data, the Win/Loss prediction model(RM2) using both the dimension reduction dividend data and the random forest algorithm was ranked highest with accuracy 56.6%.

Third, as a result of exploring and comparing the Win/Loss prediction models using the raw material mixed with the dimension reduction mixed data, the Win/Loss prediction model(M2) using the mixture and random forest algorithm was ranked highest with accuracy 57.8%.

In conclusion, the study used English Premier League game records and odds data to explore and compare the soccer Win/Loss prediction model based on the classification machine learning algorithm. It is expected that this study will be used as basic information for the use of classification machine learning algorithms according to data types when building a soccer Win/Loss prediction model by utilizing machine learning.

부 록

<부록 1> 경기기록 자료 조합별 차원축소 변인

자료	변인	중요도	자료	변인	중요도
1경기전	home team pass accuracy	0.047		home_total pass	0.032
	home_short pass	0.042		home_short pass	0.030
	home_total pass	0.041		away team tackle	0.029
	away team pass accuracy	0.041		home_key pass	0.028
	away_total pass	0.040		home team pass success	0.027
	away_short pass	0.037		away_key pass	0.025
	home team possession%	0.036	4경기전	home team possession	0.043
	away team possession%	0.035		home team corner	0.043
	home_long pass	0.028		away team possession	0.041
	home team pass success%	0.027		home_total pass	0.041
	home team shots	0.026		away_short pass	0.040
	home_key pass	0.026		away_total pass	0.038
				home team pass accuracy	0.038
2경기전	home team pass success%	0.047		home_short pass	0.036
	home_short pass	0.043		away_key pass	0.034
	home_total pass	0.041		away team openplay	0.031
	away_total pass	0.039		away team pass accuracy	0.031
	away team pass accuracy	0.037		home team pass success	0.031
	home team pass success%	0.037	5경기전	home team possession	0.047
	away_short pass	0.036		away team possession	0.045
	away team possession%	0.031		away team corner	0.045
	home team possession%	0.031		away_short pass	0.043
	away team shots on target	0.028		away team shots	0.043
	away_clearances	0.027		away team pass accuracy	0.038
	home team shots	0.026		home_total pass	0.035
				away team openplay	0.034
				away_cross	0.030
				home_short pass	0.029
				away_key pass	0.028
3경기전	away_total pass	0.059			
	away_short pass	0.056			
	away team pass accuracy	0.051			
	home team possession%	0.040			
	away team possession%	0.038			
	home team pass accuracy	0.033			

<부록 2> 배당률 자료 조합별 차원축소 변인

자료	변인	중요도	자료조합	변인	중요도
1경기전	PSH	0.051	4경기전	BWA	0.036
	BbAvH	0.049		VCA	0.033
	BbAvA	0.041		BWH	0.032
	BbMxH	0.040		IWA	0.031
	BbMxA	0.040		B365H	0.030
	PSA	0.036		BbMxD	0.030
	BWH	0.035		BbAvA	0.051
	B365H	0.035		PSH	0.049
	BWA	0.034		PSA	0.047
	WHA	0.031		VCA	0.047
	VCA	0.031		BWA	0.045
	B365A	0.030		BbAvH	0.043
2경기전	BbAvA	0.048		BbMxH	0.041
	BWA	0.043		BbMxA	0.038
	PSH	0.042		VCH	0.034
	BbAvH	0.042		IWA	0.033
	PSA	0.041		WHA	0.031
	IWA	0.037		BbAHh	0.030
	BbMxA	0.037	5경기전	BWA	0.054
	BbMxH	0.036		PSH	0.053
	B365H	0.032		BbAvA	0.050
	VCA	0.031		PSA	0.046
	BWH	0.030		IWA	0.041
	IWH	0.029		BWH	0.041
3경기전	BbAvA	0.049		BbMxH	0.040
	PSA	0.046		BbAvH	0.039
	PSH	0.044		BbMxA	0.037
	BbMxH	0.038		VCA	0.036
	BbAvH	0.036		IWH	0.035
	BbMxA	0.036		VCH	0.034

<부록 3> 혼합자료 조합별 차원축소 변인

자료	변인	중요도	자료조합	변인	중요도
1경기전	PSH	0.034	4경기전	PSH	0.023
	BbAvA	0.030		BbAvH	0.022
	BbAvH	0.028		VCA	0.020
	BWA	0.026		BWA	0.020
	B365H	0.026		BWH	0.020
	PSA	0.024		away team tackle	0.019
	BbMxH	0.023		BbAvA	0.034
	WHA	0.022		PSH	0.031
	BbMxA	0.022		BWA	0.030
	VCA	0.020		PSA	0.030
	WHH	0.020		BbAvH	0.029
	BWH	0.019		BbMxA	0.028
				BbMxH	0.028
				VCA	0.027
2경기전	PSH	0.028	5경기전	BbAvA	0.032
	BbAvH	0.028		BbAvH	0.031
	BbMxH	0.028		PSH	0.031
	PSA	0.026		PSA	0.030
	BbAvA	0.025		BWA	0.030
	BWA	0.025		BWH	0.026
	BbMxA	0.023		IWH	0.026
	BWH	0.022		BbMxH	0.025
	IWA	0.022		BbMxA	0.024
	VCA	0.022		B365H	0.024
	B365H	0.021		away_total pass	0.021
	B365A	0.019		away_short pass	0.021
3경기전	away_total pass	0.031			
	BbAvA	0.026			
	PSA	0.024			
	BbMxA	0.024			
	BbMxH	0.023			
	away_short pass	0.023			