ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2022 March: 28(1): 263~286 http://dx.doi.org/10.13088/jiis.2022.28.1.263

미국 프로농구(NBA)의 플레이오프 진출에 영향을 미치는 주요 변수 예측: 3점과 턴오버 속성을 중심으로

안세환
한양대학교 기술경영학과
(hwan86@hanyang.ac.kr)

김영민
한양대학교 기술경영학과
(yngmnkim@hanyang.ac.kr)

본 연구는 웹 크롤링을 이용하여 1990년부터 2022년까지 총 32개년에 해당하는 NBA 통계 정보를 획득하고, 탐색적 데이터 분석을 통해 관심 변수를 관찰하고 관련된 파생변수를 생성한다. 입력 데이터에 대한 정제 과정을 거쳐 무의미한 변수들을 제거하고, 남은 변수에 대한 상관관계 분석, t 검정 및 분산분석을 수행하였다. 관심 변수에 대해 플레이오프 진출/미진출 그룹 간 평균의 차이를 검정하였고, 이를 보완하기 위해 순위를 기준으로 하는 3개 집단(상위/중위/하위) 간 평균 차이를 재확인하였다. 입력 데이터 중 올해 시즌 데이터만을 테스트 세트로 활용하였고, 모델 훈련을 위해서는 훈련 세트와 검증 세트를 분할하여 5-fold 교차검증을 수행하였다. 교차검증 결과와 시험 세트를 이용한 최종 분석 결과를 비교하여 성능 지표에서 차이가 없음을 확인함으로써 과적합 문제를 해결하였다. 원시 데이터의 품질 수준이 높고, 통계적 가정을 만족하기 때문에 적은 수준의 데이터 세트임에도 불구하고 대부분 모델에서 좋은 결과를 나타냈다. 본 연구는 단순히 머신러닝을 이용하여 NBA의 경기 결과를 예측하거나 플레이오프 진출 여부만을 분류하는 것에서 그치지 않고, 입력 특성의 중요도를 파악하여 높은 중요도를 갖는 주요 변수에 본 연구의 관심 대상 변수가 포함되는지를 확인하였다. Shap value의 시각화를 통해 특성 중요도의 결과만으로 해석할 수 없었던 한계를 극복하고, 변수의 진입/제거 과정에서 중요도 산출에 일관성이 부족하다는 점을 보완할 수 있었다. 본 연구에서 관심 대상으로 분류했던 3점 및 실책과 관련된 다수의 변수가 미국 프로농구에서의 플레이오프 진출에 영향을 미치는 주요 변수에 포함되는 것으로 나타났다. 본 연구는 기존의 스포츠 데이터 분석 분야에서 다루었던 경기 결과, 플레이오프 및 우승 예측 등의 주제를 포함하고 분석을 위해 여러 머신러닝 모델을 비교 분석했다는 점에서 유사성이 있지만, 사전에 관심 속성을 설정하고, 이를 통계적으로 검증함으로써 머신러닝 분석 결과와 비교하였다는 측면에서 차이가 있다. 또한 XAI 모델 중 하나인 SHAP를 이용하여 설명 가능한 시각화 결과를 제시함으로써 기존 연구와 차별화하였다.

주제어 : 미국 프로농구 플레이오프 분석, 기계학습, 통계적 분석, 샤플리, 특성 중요도

논문접수일 : 2022년 3월 4일 논문수정일 : 2022년 3월 12일 게재확정일 : 2022년 3월 22일
원고유형 : 급행논문 교신저자 : 김영민

1. 서론

미국 프로농구 (National Basketball Association, NBA)는 30개 팀으로 구성된 세계에서 가장 인기 있는 스포츠 리그 중 하나로 세계 최고의 남자 프로 농구 리그로 평가받는다. 많은 스포츠 팬들이 NBA 경기, 특히 결승전을 결정짓는 플레이오프를 관람하고 있으며, 약 70년의 NBA 역사는 모든 팀에 대한 풍부한 통계를 제공하고 있다. 각 통계는 경기당 득점, 리바운드 수 및 어시스트 수와 같은 팀과 선수의 성과 평가를 제공한다. 이러한 통계 데이터를 수집하여 경기 또는 전체 시즌에 대한 예측을 할 수 있다.

스포츠 경기를 모델링하고 예측하는 것은 스

포츠 팀이 승리하기 위한 전략을 더 잘 이해하고 적용하는 데 도움이 될 뿐만 아니라 팬들이 스포츠 경기를 더 잘 즐길 수 있도록 도움을 준다(Nunes and Sousa, 2006). 또한 스포츠 경기에 대한 승부 예측은 분석가에게는 경기의 흐름에 대한 정보를 제공할 수 있으며, 경기의 승리와 패배에 영향을 미치는 변인을 도출하여 팀의 전략 수립을 가능하게 한다(Lee et al., 2020).

많은 선행연구들에서(Cheng et al., 2016; Kaur amd Jain, 2017; Thabtah et al., 2019; Migliorati, 2020; Chen et al., 2021; Wang et al., 2022) NBA 경기 결과 예측 및 주요 변수 도출을 위한 다양한 종류의 머신러닝 분석 기법이 제안되었다.

21세기 이전까지 3점 시도는 점수를 얻는 효율이 낮은 방식으로 여겨졌으나, 경기 전술이 풍부해지면서 NBA 모든 팀은 더 나은 공격 성과를 얻기 위해 더 많은 3점을 시도하였다(Geng and Hu, 2020).

NBA뿐만 아니라 세계 농구는 3점이 각광 받는 시대이다. 2010년대 초반까지만 해도 안정적인 골밑 득점이 가장 좋은 공격 방법으로 평가받았지만 불과 7-8년 사이에 흐름이 완전히 바뀌었다. Stephen Curry (Golden State Warriors), James Harden (Philadelphia 76ers) 등 압도적인 슈팅 능력을 가진 선수들이 대세로 떠올랐으며, 이러한 흐름에 따라서 스페인, 러시아, 이탈리아, 프랑스 등 유럽 명문리그에서도 3점의 중요성이 높아졌다.

신장이 2m 10cm(약 7피트)에 달하는 빅맨에게도 3점 슈팅의 구사 능력은 필수가 되었다. Anthony Davis (211cm·Los Angeles Lakers), Karl-Anthony Towns (216cm·Minnesota Timberwolves), Joel Embiid (213cm·Philadelphia 76ers) 등 2m 10cm가 넘는 리그 정상급 센터들도 외곽에서 3점을 시도하며, 속공 상황에서 조차 레이업 슈팅이 아닌 3점을 시도하는 장면을 어렵지 않게 볼 수 있다. 유럽에서도 골밑에만 위치하는 정통 빅맨이 점차 줄어들고 있고, 변화에 둔감했던 한국 농구도 빅맨을 이용한 센터 농구에서 탈피해 3점 시도가 늘어나는 추세다. 이들의 3점은 아예 팀의 주요 옵션 중 하나로 자리매김하고 있으며, 기본적으로 3점 시도 자체가 급격하게 증가했다.

이처럼 상대 수비를 밖으로 끌어내 공격할 공간을 넓히고 한 번의 공격으로 3점을 얻을 수 있다는 효과에 주목하기 시작하면서 3점은 현대 농구의 가장 효과적이고 효율적인 공격 방법으로 자리 잡았다. 막강한 공격력을 보유한 팀이 리그를 지배하면서 3점이 약한 팀은 우승할 수 없는 시대가 되었다.

Kaur and Jain (2017)의 연구에 따르면 3점 성공률(3P%)이외에도 경기 결과에 영향을 미치는 주요 속성으로 실책(TOV)이 포함되는 분석 결과를 제시하고 있다. 실제로 현대 농구에서의 실책의 중요성이 나날이 높아지고 있다. 실책으로 이어진 상대 팀의 속공 플레이로 인한 손쉬운 득점은 곧 팀의 실점으로 귀결된다. 따라서 실책으로 인한 실점은 결정적인 승부처에서 경기 결과를 뒤바꿀 수 있는 매우 중요한 요인이다.

Kim et al. (2017)의 연구는 한국 프로농구의 승패 결정요인을 추정하기 위해 공격변인과 수비변인으로 구분하여 로지스틱 회귀 분석 결과를 제시하였다. 분석 결과에 따르면 공격변인 중 2점 성공률(2P%), 3점 성공률(3P%), 실책(TOV)이 통계적으로 유의한 것을 확인하였다.

본 연구에서는 약 30년간(1990-2022) 미국 프로농구 시즌 통계 정보인 컨퍼런스 순위(Conference Standings)와 경기별 통계(Per Game Stats) 데이터를 활용한다. 해당 테이블에서 추출

되지 않은 유효 필드 골 비율(eFG%)과 실책율(TOV%) 등의 수치는 기존 데이터를 활용하여 부가적으로 산출한다. 유효 필드 골 비율과 실책율 요인은 단순하게 수치의 양적 측면 외에도 질적 측면을 고려한다는 점에서 프로농구 경기 결과를 분석하는 다양한 연구에서 주요 속성으로 활용되고 있다(Kubatko et al., 2007; Mandićet al., 2019; Liu, 2021).

특히 플레이오프 진출 시 영향을 미치는 주요 변수를 파악하기 위해 3점과 실책 속성을 중심으로 데이터를 탐색하고 이와 관련된 파생변수 생성이 필요하다. 따라서 유효 필드 골 비율 지표의 경우 동일한 개체 수가 포함될 수 있도록 그룹별(상, 중, 하) 범주화하였고, 리그 평균 유효 필드 골 비율을 산출하여 각 팀별로 해당 비율이 평균을 초과하거나 그렇지 못한 집단을 구분하였다.

이어서 플레이오프 진출 집단과 미진출 집단 간에 3점 및 실책과 관련된 속성에 있어서 평균에 차이가 있는지 독립표본 t검정을 수행하고, 이를 바탕으로 여러 머신러닝 알고리즘을 적용하여 플레이오프 진출 여부에 대해 예측한다.

마지막으로 머신러닝 분석을 통해 도출된 플레이오프 진출에 영향을 미치는 주요 속성별 중요도를 XAI (Explainable AI) 모델 중 하나인 SHAP (SHapley Additive exPlanation)을 통해 확인하고 선정했던 속성을 중심으로 그 결과를 평가한다.

본 연구의 구성은 다음과 같다. 2장에서 스포츠 데이터 분석과 관련된 연구, 특히 NBA의 경기 결과 예측을 위해 머신러닝 알고리즘을 이용한 관련 연구를 소개한다. 3장에서는 본 논문의 연구방법을 소개하고, 연구에 활용된 통계 및 머신러닝 개념을 기술한다. 4장에서는 모형별 예측 결과를 분석하고 평가한다. 마지막 5장에서는 연구 결과를 정리하고 시사점 및 추후 연구방향에 대해 논의한다.

2. 관련 연구

2.1. 스포츠 데이터 분석

과거 스포츠 데이터 분석 분야는 데이터 수집의 한계로 통계 분석 기법을 이용하여 모집단에 대해 추정하거나 기초 통계량(평균, 분산, 엔트로피, 최대·최소 값 등)과 같은 데이터 세트의 일부 통계적 특성을 개략적으로 분석하는 연구에 국한되었다(Bai and Bai, 2021).

스포츠 관람이 활성화되고 사람들의 관심이 증가함에 따라 스포츠 산업은 지난 수십 년 동안 상당한 성장을 이루었다. 단순하게 스포츠를 보고 즐기는 것을 넘어서 미국과 유럽 등 세계 전역에서 스포츠 경기에 대한 배팅이 합법적으로 운영되고 있다.

이러한 이유로 다양한 방식의 스포츠 예측 모델이 베팅 회사가 베팅 확률 산출하고, 경기 결과를 예측하기 위한 목적으로 주로 활용되고 있다(Yazbek et al., 2021). 반면에 배팅을 하는 bettor (내기하는 사람)에게 결과 예측 서비스를 제공하는 영리 목적의 스포츠 데이터 분석 업체에서 다양한 방식으로 예측 모델을 연구·개발하고 있다.

통계 분석 기술은 스포츠 데이터 연구에서 중요한 역할을 해왔지만, 스포츠 산업과 빅데이터 기술의 발달로 인해 머신러닝, 데이터 마이닝, 딥러닝 등의 기술이 스포츠 빅데이터 연구에 활용되고 있다(Al-Jarrah et al., 2021).

〈Table 1〉 Summary of Studies on ML in Basketball

#	Data set	Aim	Type of ML models
Albert et al. (2022)	Data of a single person's single-season statistics about 17,000 players	To improve the accuracy of the predictions, for predicting NBA All-Stars	SVC, KNN, Decision tree, Gaussian process, Random forest, AdaBoost, Multilayer perceptron
Chen et al. (2021)	Data for every single NBA game in the 2018–2019 season	To verify the hypothesis that the highest level of European basketball is becoming quantitatively and qualitatively more similar to the NBA	ELM, MARS, XGBoost, SGB and KNN
Geng et al. (2020)	Advanced statistics data from 1984 to 2018	To predict the results of the entire NBA playoffs using one trained model	Genetic programming algorithm
Horvat and Job (2019)	NBA seasons 2009–2017	To predict the outcome of NBA games Naïve ML algorithm	Naïve ML algorithm
Mandić et al. (2019)	Box-score data for all NBA and Euroleague games in the period 2000-2017	To propose a hybrid data-mining-based scheme for predicting the final score of an NBA game	Statistical analysis
Hsu et al. (2018)	Data per game for player attributes from 2011 to 2018 season	To predict team rankings instead of classifying their rankings based on regression method	Support vector machine, Polynomial regression, Random forest
Lam (2018)	NBA seasons 2013–2014	To present modeling approach for one-match-ahead forecasting in two team sports featured with winning probability calculation	Bayesian regression
Bianchi et al. (2017)	Data from 82 games of 2015–2016 NBA regular season	To describe new roles of players during the game	Neural networks (self-organizing maps, fuzzy clustering)
Leicht et al. (2017)	Women's basketball matches during the 2004–2016 Olympic Games	To examine the relationship between team performance indicators and match outcome during the women's basketball tournament	Binary logistic regression, conditional interference classification tree
Pai et al. (2017)	NBA seasons 2008–2010	To investigate developed a HSVMDT model for analyzing the game's outcomes in the NBA	Support vector machine, Decision tree
Cheng et al. (2016)	Data from 14 basic technical features from 2007 to 2015 season	To predict the outcome of NBA matches	Maximum entropy model, k-means clustering, Naïve Bayes, Logistic regression, BP neural networks, Random forest
Kempe et al. (2015)	Player's positional data during matches	To compare its performance to the common dynamical controlled network approach to analyze team sport position data	Neural networks
Lopez and Matthews (2015)	Predictions(data) for NCAA basketball tournaments	To analyze the accuracy of traditional methods vs cutting-edge predictive algorithms	Logistic regression, Log-loss function
Zimmermann et al. (2013)	NCAAB seasons 2008–2013	To explore the use of ML techniques for making NCAAB match outcome prediction	Random forest, Naïve Bayes, Multilayer perceptron neural
Schmidt (2012)	Data from 21 participants that performed 20 free-throw trials	To analyze the movement patterns of free-throw shooters at different skill levels	Neural networks

축구 경기 결과 예측을 위해 베이지안 네트워크 모델에 기반한 머신러닝 분석(Joseph et al., 2006; Rahman et al., 2018)에 관한 연구가 수행되었고, Prasetio and Harlili (2016)는 영국 프리미어 리그 (English Premier League, EPL)의 시즌 데이터를 로지스틱 회귀 분석 모델로 분석하여 경기 결과를 예측하였다.

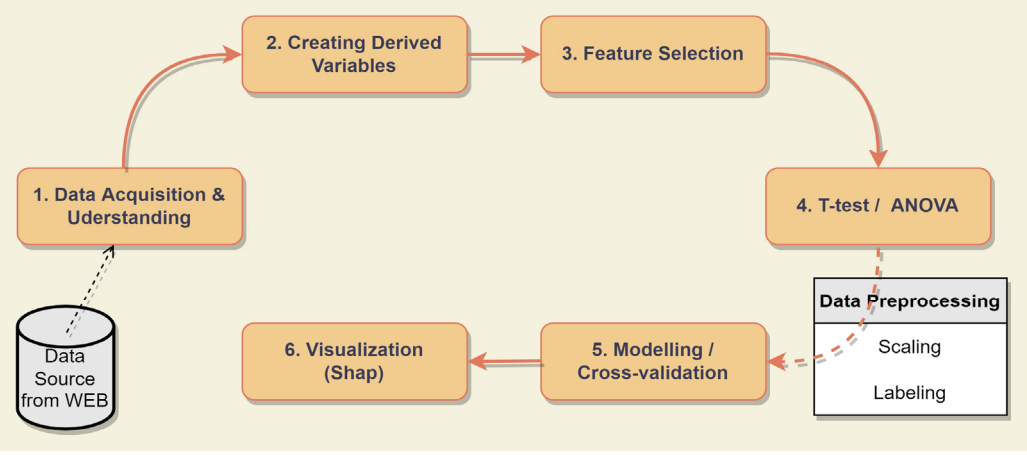
MLB 정규시즌 경기의 승패를 k-NN, SVM, Decision Tree 등의 알고리즘을 통해 예측한 (Soto, 2016) 야구의 사례와 대학 미식축구 경기의 승자 예측을 두 경쟁 팀의 통계를 비교하는 대신 과거의 경기 결과를 기반으로 하는 데이터 마이닝 접근 방식을 적용한 연구가 제안되었다 (Leung et al., 2014).

그 밖에도 스포츠 분석과 관련된 연구가 매우 다양한 방식으로 수행되었으며, Lai et al. (2018)의 연구는 탁구 선수의 성공 예측을 위해 시합의 네트워크 기여도를 추정함으로써 위상적 속성이 예측력을 높이는데 효과적인 역할을 한다는 것을 확인하였다.

2.2. 농구에서 머신러닝 적용 연구

팀 단위 스포츠인 축구, 농구, 야구 등을 비롯하여 개인 단위의 스포츠인 테니스, 골프, 사이클 등의 다양한 스포츠가 있지만, 특히 NBA는 세계에서 가장 많이 시청되는 스포츠 중 하나로 많은 팬들에게 다가오는 경기의 결과를 기대하게 한다(Chen et al., 2019).

<Table 1>은 농구 분야에서 머신러닝을 적용



〈Figure 1〉 Procedure of the Proposed Research Method

한 연구를 시간순으로 나열한 것이다. 2010년대부터 최근 2020년대까지의 농구, 특히 NBA와 관련된 연구를 포함하며, 일부 미국 대학 농구와 여성 농구 경기에 관한 연구도 소개되었다.

대부분의 선행연구는 기본적인 통계 분석을 포함하여 서포트 벡터 머신, 로지스틱 회귀, 랜덤 포레스트, 신경망 등 여러 머신러닝 기법을 활용해 이루어졌고, 주로 경기 결과에 대한 예측을 연구목적으로 수행되었지만, 그 밖에도 우승팀 예측, 순위 예측, NBA 올스타 선발 예측, 경기 승리에 대한 선수 기여도 평가, 위치 데이터 기반의 전술 분류 등 다양한 주제를 포함하였다 (Zimmermann et al., 2013; Lopez and Matthews, 2015; Cheng et al., 2016; Hsu et al., 2018; Lam, 2018; Pai et al., 2017; Albert et al., 2022;). Horvat and Job (2019)의 연구는 농구에서 최종 예측 결과를 비교하는 것 외에도 속성 선택 및 추출 방법의 활용에 대한 통찰을 제시하고 있다.

본 연구는 관련 연구들에서 머신러닝 분석을 통해 예측 결과를 도출하는 기존의 방식에서 벗어나 사전에 관심 변수를 설정하고, 해당 변수에 대한 활용도를 높이기 위해 초기 변수 이외에도 파생변수를 부가적으로 생성하였다. 종속변수의 범주(집단)에 따른 관심 속성과의 통계적 검정 과정을 거쳐 실제 머신러닝 분석으로부터 도출된 주요 독립변수와 비교 분석하였고, 그 결과에 대한 해석력을 높이기 위해 설명 가능한 인공지능(XAI) 기법을 적용하여 재확인하였다. 기존의 연구에서 분석 결과에 대한 해석을 위해 주로 앙상블 모형을 이용하여 변수가 가지는 가중치 및 이득 등에 초점을 두어 그 영향력을 순위로 산출하는 것에 국한되었다. 본 연구는 SHAP을 이용하여 각 속성별 영향력뿐만 아니라 '어떻게' 영향을 주고 있는가에 대한 방향성을 제시함으로써, 특성별 중요도와 관리 방향을 동시에 활용하여 인사이트를 도출하고, 이를 실무적 활용 가능하다는 측면에서 연구적 의의를 찾아볼 수 있다.

3. 연구 방법

본 연구에서 제안하는 연구 방법의 절차는

<Figure 1>과 같다. 웹 크롤링 (Web crawling)을 이용하여 NBA 통계 정보를 획득하고, 탐색적 데이터 분석을 통해 추가적인 파생변수를 생성한다.

본 연구에서는 팀별 플레이오프 진출에 영향을 미치는 요인으로서 3점 시도 (3PA), 성공 횟수 (3PM) 및 성공률 (3P%)과 실책 (TOV)에 대한 속성과 관련 속성으로부터 구해진 리그 평균 3점 성공률 (3P%_Mean), 리그 평균 3점 성공률 초과 여부 (3P_OverMean) 등의 파생변수에 중점을 두어 분석을 진행한다.

부가적으로 Oliver (2004)에 의해 제안된 '승리의 4가지 요소' 중 '3점'과 '실책' 속성과 관련되며, 본 연구의 핵심 요소에 해당하는 유효 필드골 비율 (eFG%) 및 실책율 (TOV%) 변수를 관심변수로 포함한다.

3.1. 관심 속성 검정

입력 데이터 중 범주형에 해당하며, 고유값 (nunique)의 수가 많아서 인스턴스를 식별하는 것 이외에 의미가 없는 경우와 미사용 변수에 대해 1차 변수 선정 과정을 거쳐서 제거한다.

상관관계 분석, t 검정 (t-test) 및 분산분석 (Analysis of Variance, ANOVA)을 수행함으로써 변수 간 관계를 파악하고, 예측변수인 플레이오프 진출/탈락 집단 간에 관심 속성을 중심으로 집단 간 평균에 차이가 있는지 검정한다.

본 연구에서는 플레이오프 진출/탈락 그룹을 실험군으로 하여 '3점'과 '실책' 속성과 관련된 요인들을 대상으로 두 집단 간 평균의 차이 ($\overline{x}_1 - \overline{x}_2$)를 두 군의 개체를 통합할 때 개체 간 변동의 차이, 즉 표준오차 (Standard Error of Mean, SEM)로 나누어 구한다. 산출된 결과에 따라 검정통계량이 유의하면 귀무가설을 기각하고, 두 실험군의 평균이 유의하게 차이가 있다고 간주할 수 있으며, 다음의 식 (1)과 같다.

$$t = \frac{\overline{x}_1 - \overline{x}_2}{SEM_{Diff}} \quad (1)$$

동일한 방식으로 순위를 기준으로 3 수준(상, 중, 하)으로 그룹화한 'Lev_Rk'를 '3점'과 '실책' 속성과 관련된 요인을 중심으로 평균 차이를 검정하기 위해서 분산분석을 사용한다. 일원배치 분산분석 (one-way ANOVA)에 대한 실험군이 n개인 i번째 처리군에서 j번째 관찰치의 총 변동 (Total Sum of Square, SST)은 처리군 간의 변동 (Between Sum of Square, SSB)과 처리군 내 개체 간의 변동인 오차제곱합 (Error Sum of Square, SSE)으로 분해할 수 있고, 다음의 식 (2)와 같다.

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \overline{\overline{x}})^2$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}(\overline{x}_{i.} - \overline{\overline{x}})^2 + \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \overline{x}_{i.})^2 \quad (2)$$

3.2. 모형 적용

모델 구현에 앞서 수치형 데이터의 경우 표준화를 통한 스케일링을, 범주형의 경우 데이터 특성에 맞게 각각 바이너리 및 레이블 인코딩을 수행함으로써 입력 데이터에 대한 전처리를 수행한다.

분석 모델의 선정은 기존 스포츠 경기 결과 예측에 관한 선행연구에서 빈번하게 활용되고, 우수한 성능을 보였던 모델을 선정한다. 본 연구에

서는 로지스틱 회귀 (Cheng et al., 2016; Kaur amd Jain, 2017; Horvat et al., 2020), 랜덤 포레스트 (Cheng et al., 2016; Horvat et al., 2020; Young et al., 2020; Jain et al., 2021; Albert et al., 2022), Adaboost (Albert et al., 2022), XGboost (Young et al., 2020)의 4가지 모델을 활용하여 그 결과를 비교한다.

특히 랜덤 포레스트와 XGboost 등 앙상블 모델을 통해 도출된 특성 중요도 (Feature importance)와 SHAP을 통해 관심 속성의 중요도를 비교·분석하고, 변수별 영향력을 해석한다.

3.2.1. **로지스틱 회귀분석** (Logistic Regression)

로지스틱 회귀 알고리즘은 예측뿐만 아니라 분류 문제에서도 사용할 수 있으며, 특히 클래스가 0, 1로 구분된 이진 분류 문제에 널리 활용되고 있다.

로지스틱 회귀 모형은 일반적인 선형 회귀 방식과 같이 입력 속성의 가중치 총합에 편향을 더함으로써 식 (3)과 같이 결과값의 로지스틱을 출력한다.

$$\hat{p} = h_\theta(X) = \sigma(\theta^T X) \tag{3}$$

로지스틱의 0과 1사이의 값을 출력하는 S자 형태의 시그모이드 함수 (Sigmoid function)는 식 (4)와 같고, 식 (5)와 같이 특정 개체가 각 클래스에 포함될 확률을 추정함으로써 예측값을 산출할 수 있다.

$$\sigma(t) = \frac{1}{1 + e^{(-t)}} \tag{4}$$

$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0.5, \\ 1 & \text{if } \hat{p} \geq 0.5 \end{cases} \tag{5}$$

3.2.2. **랜덤 포레스트** (Random Forest)

앙상블 학습 (Ensemble learning)이란 학습 알고리즘을 개별적으로 사용하는 것보다 더 좋은 예측 성능을 얻기 위해 다수의 학습 알고리즘을 사용하는 방식으로 의사결정나무의 부정확성을 보완할 수 있는 알고리즘이다.

랜덤 포레스트는 분류 및 회귀 등에 사용되는 앙상블 학습 방식의 일종으로 훈련 세트로부터 무작위의 서브 세트를 만들어 다수의 결정 트리 분류기를 훈련할 수 있다. 훈련 세트에서 중복을 허용하여 샘플링하는 배깅 (Bagging) 방식의 경우 하나의 훈련 세트를 동일한 분류기에서 반복 사용할 수 있다.

모든 의사결정 트리가 생성되고 학습이 종료되면, 앙상블은 모든 예측값을 통합하여 새로운 샘플에 대한 예측을 각 트리에서 투표하여 집계함으로써 예측 성능을 높일 수 있다.

랜덤 포레스트의 또 다른 장점으로 분류 또는 회귀 문제에서 각 특성에 대한 상대적 중요도 산출이 가능하다. 변수의 중요도를 측정하기 위해 데이터 세트에 맞춰 알고리즘을 훈련하는 과정에서 각 데이터 샘플에 대한 oob (out of bag) 오차는 데이터의 실제값과 예측값 사이의 오차로 정의한다. 각 속성의 중요도 점수는 oob 오차와 치환된 데이터 세트의 oob 오차 간 차이의 평균으로 정의되고, 중요도가 높은 변수에는 낮은 변수보다 높은 점수를 할당한다.

3.2.3. Adaboost

부스팅 (Boosting) 방식의 일종으로 Adaboost

는 이전 모형에서 과소적합했던 학습 샘플의 가중치를 높여 이를 보완하는 방식으로 학습한다. 학습에 이용된 개별 분류기에 서로 다른 가중치를 반영하여 최종 분류기를 생성하는 부스팅의 발전적 형태라고 할 수 있다.

랜덤 포레스트와 기본적인 아이디어는 유사하지만 랜덤 포레스트가 결정 트리에 미리 정해진 크기가 없는 반면, Adaboost에서 트리는 대부분 결정 노드 하나와 리프 (Leaf) 노드 두 개로 구성하며, 이러한 의사결정나무의 배열을 stump이라고 한다. 또한 랜덤 포레스트의 분류기가 학습 시 서로 독립적으로 생성되는 것과 달리 Adaboost는 순서를 가지는 일련의 분류기에 대해 잘못 분류된 훈련 샘플의 가중치를 상대적으로 높여 업데이트함으로써 새로운 분류기 생성에 영향을 미친다. 각 샘플의 초기 가중치 w_i를 $\frac{1}{m}$로 설정하고 처음 분류기가 학습하면, 가중치가 적용된 오류율 ϵ_m이 훈련 세트에 대해 산출되는 과정은 다음의 식 (6)과 같다.

$$\epsilon_m = \frac{\sum_{y_i \neq k_m(x_i)} w_i^{(m)}}{\sum_{i=1}^{N} w_i^{(m)}} \tag{6}$$

분류기 가중치에 대한 α_m은 식 (7)과 같이 산출한다. η은 학습률을 의미하며, 오류율 ϵ_m이 0.5에 가까워지면 가중치가 0에 근사하고, 0.5보다 높으면 음수의 가중치 값을 갖는다.

$$\alpha_m = \eta \frac{1}{2} ln \frac{1-\epsilon_m}{\epsilon_m} \tag{7}$$

이어서 잘못 분류된 샘플의 가중치가 커지도록 업데이트가 되면 모든 샘플의 가중치에 대해 정규화 과정을 거친다. 최종적으로 새 분류기에서 업데이트된 가중치를 이용하여 훈련하고, 이러한 과정을 반복하다가 지정한 분류기 수에 도달하거나 높은 성능의 분류기가 만들어지면 과정을 중단한다.

3.2.4. XGboost

부스팅 (Boosting) 방식을 이용하여 구현한 알고리즘 중 하나인 GBM (Gradient Boosting Method) 알고리즘을 병렬 학습이 가능하도록 보완한 모델이 XGboost이다. 분류와 회귀의 모든 문제에 활용 가능하며, GBM과 달리 분류회귀트리 (CART) 기반의 트리를 사용하며, 다양한 손실함수 (Loss function)에 대한 지원과 튜닝을 통해 모형 복잡도에 대한 고려도 가능하다는 장점이 있다. 각 CART 모델을 훈련하기 위한 손실함수는 다음의 식 (8)과 같다. 여기서 l은 예측값 $\hat{y}_i$과 실제값 y_i사이의 차이에 대한 손실함수를 나타내고, 정규화 항인 Ω는 모델의 복잡도에 페널티를 주어 과적합 (Overfitting) 방지하고 예측력을 높인다. 여기서 T는 약한 분류기로 사용하는 트리의 leaf 노드 수가 되고 w_j는 각 leaf 노드의 점수를 의미한다.

$$\begin{aligned} L(\phi) &= \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \\ where\ \Omega(f) &= \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \end{aligned} \tag{8}$$

학습 결과에서 발생한 손실을 또 다른 분류기를 통해 학습하여 손실을 줄이고, 이러한 과정을

반복하며 손실이 최소화되도록 하는 split point를 찾는 것이 XGboost 모델의 학습 방향이다. 즉, 트리가 가지를 나누는 시점에서 얻어지는 이득에 대해 계산하여 해당 점수가 음수일 때 가지치기를 반복하면 최종적으로 점수가 높은 트리가 조합된 최적의 분류 모델을 얻을 수 있다.

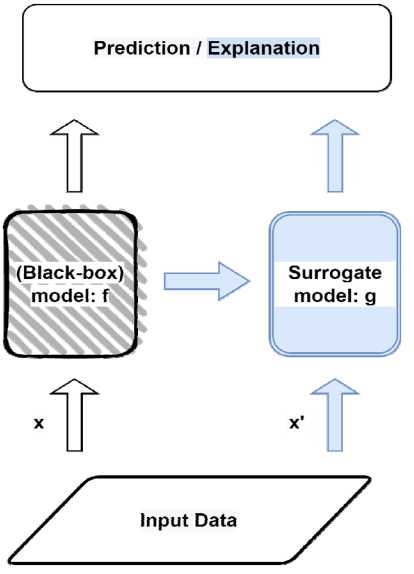
3.3. Explainable AI by SHAP

XAI는 복잡한 딥러닝 및 머신러닝의 분석 결과를 비전문가도 쉽게 이해하고 사용할 수 있는 기술로써 대표적으로 LIME, SHAP, ELI5, Interpretml 등과 같은 다양한 라이브러리가 소개되고 있다.

본 연구에서는 SHAP 모형을 통해 중점적으로 다루었던 입력 속성들이 실제 어느 정도의 중요도를 가지고 어떠한 방향으로 예측 결과에 영향을 보이는지 확인한다.

SHAP의 기계학습 모델에 대한 결과 해석은 협력적 게임 이론에 기반한 최적의 Shapley value를 사용하여 정의될 수 있다. 예를 들어, 특정 입력 속성에 대한 중요도를 파악하기 위해 입력 속성 간에 조합을 만들어 해당 속성의 존재에 따른 평균 예측값의 변화를 통해 값을 도출한다. 이는 예측값에 대한 입력 속성의 기여도를 의미한다. Shapley value는 지역적 (Local) 및 전역적 (Global) 해석이 모두 가능하므로 대조 설명 (Contrastive explanations)이 불가능한 LIME과 같은 지역적 모델과 차별화될 수 있다.

SHAP의 목적은 예측한 결과에 대한 각 특성의 기여도를 계산하여 개별 테스트 데이터의 예측값을 설명하는 것이 핵심이다. 즉, 블랙 박스 모델 f를 통해 예측된 결과를 간소하게 변형한 x'를 입력으로 하여 대리의 모델 (Surrogate Model)인 g를 찾고, 실험 데이터에 대한 해석을 가능하게 한다. 다음 <Figure 2>는 SHAP의 작동 원리를 도식화하였다.



〈Figure 2〉 Understanding of SHAP flow

4. 실험 및 결과

NBA 플레이오프 진출 예측 모델의 실험환경으로는 Window 10 운영체제와 Python 3.8 프로그래밍 언어를 사용하였으며, 활용 프레임워크와 라이브러리는 Pandas, Numpy, Scikit-Learn, BeautifulSoup, 그리고 시각화를 위한 seaborn 및 shap 등이다.

4.1 데이터 획득 및 탐색

본 논문에서 사용된 실험 데이터는 NBA 팀별

〈Table 2〉 Summary of studies on ML in basketball

Year	Season	**2PM**	2 Point Field Goals Made
Team	Team Name	**2PA**	2 Point Field Goals Attempted
W	Wins	**2P%**	2 Point Field Goals %
L	Losses	**eFG%**	Effective Field Goals %
W/L%	Win %	**eFG%_Mean**	League Average Of eFG %
Rk	Rank	**eFG_OverMean**	Above eFG%_Mean
Lev_Rk	Level Of Rank	**FTM**	Free Throws Made
PS/G	Points Per Game	**FTA**	Free Throws Attempted
PA/G	Opponent Points Per Game	**FT%**	Free Throw %
SRS	Simple Rating System	**ORB**	Offensive Rebounds
GP	Games Played	**DRB**	Defensive Rebounds
MP	Minutes Played	**TRB**	Total Rebounds
FGM	Field Goals Made	**AST**	Assists
FGA	Field Goals Attempted	**TOV**	Turnovers
FG%	Field Goal %	**TOV%**	Turnover %
3PM	3 Point Field Goals Made	**STL**	Steals
3PA	3 Point Field Goals Attempted	**BLK**	Blocks
3P%	3 Point Field Goals %	**PF**	Personal Fouls
3P%_Mean	League Average Of 3P %	**Playoffs**	Advance To Playoffs
3P_OverMean	Above 3P%_Mean		

통계 데이터를 온라인 웹사이트에서 추출하였다. 수집 데이터의 세부 항목은 NBA 컨퍼런스 순위 (Conference Standings)와 경기별 통계 (Per Game Stats)의 테이블 정보에 해당하고, 수집 데이터의 처리 과정은 다음 방식을 따른다.

1) 웹 데이터 크롤링
   (https://www.basketball-reference.com/)
2) 테이블 간 병합 및 컬럼 수정
3) 최종 데이터의 품질 확인 및 테이블 수정
4) 데이터 탐색 및 파생변수 생성

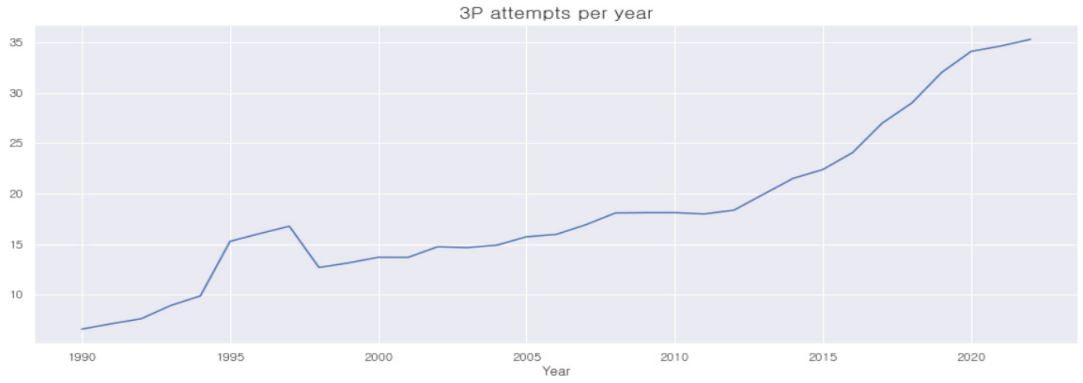
실험에 이용한 데이터 세트는 NBA 팀별 32개 시즌(1990/1991 ~ 2021/2022) 통계 데이터로 인스턴스 963건과 컬럼 39개로 구성되었다. 입력 컬럼에 대한 변수 정보는 다음의 <Table 2>와 같다.

본 논문에서는 NBA 플레이오프 진출에 영향을 미치는 변수로 3P 및 TOV와 관련된 속성을 중심으로 데이터를 탐색하였다. 특히 3P과 관련된 속성 중 연간 평균 3점 시도 횟수가 꾸준하게 증가하는 패턴을 나타내고 있다.

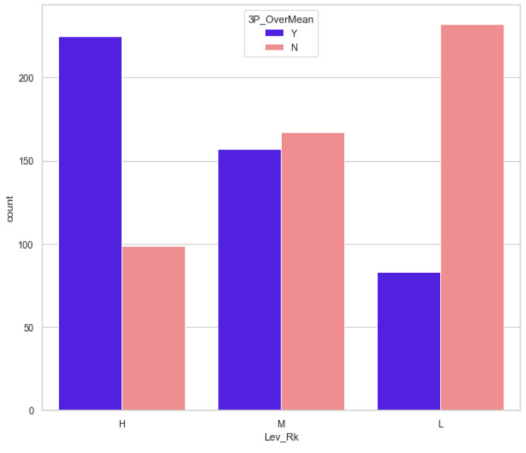
다음의 <Figure 3>을 보면 1990년부터 2022년까지 3점 시도 횟수가 점진적으로 증가하는 추세를 확인할 수 있다. 1995년부터 1998년까지 특정 구간에서 시도 횟수가 급격하게 늘어난 것은 3시즌 동안 3점 라인이 7.24m에서 6.71m로 일시적으로 줄었던 요인 때문이다.

3P과 관련된 속성 중 리그 평균 3점 성공률을 기준으로 평균 이상인 그룹과 그렇지 않은 그룹에 대한 속성과 리그 승률 기준(상/중/하) 그룹에 대한 속성을 추가하여 살펴본 빈도 플롯은 다음의 <Figure 4>와 같다.

실제로 승률이 높은 구간에 속할 때, 리그 평균 3점 성공률을 상회하는 경우가 하위 그룹에 비해 월등하게 높은 것으로 나타났다. 중위권 그룹의 경우 평균 3점 성공률에 대한 개체 수에 큰 차이가 없었다.

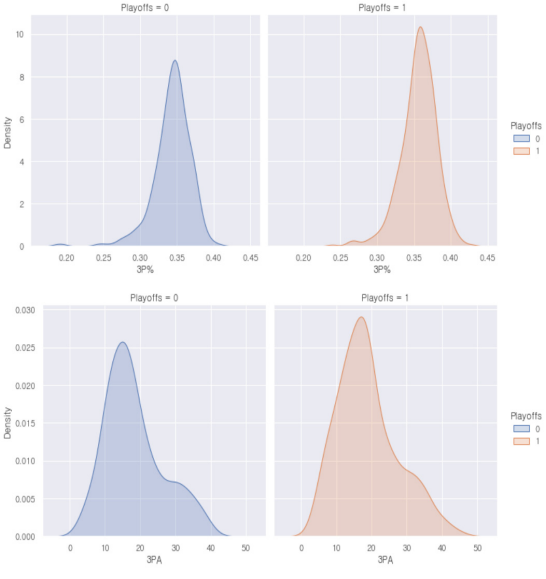


〈Figure 3〉 Trend of 3-point attempts



〈Figure 4〉 Comparison of 3-point Success rate by Ranking Group

마찬가지로 <Figure 5>의 distplot 결과를 보면, 플레이오프 진출과 탈락 그룹에 따라서 3점 성공률과 시도 횟수에 차이를 확인할 수 있다. 예를 들어, 상단의 플롯은 플레이오프 진출 여부에 따른 '3P%'의 분포 구간을 나타내고 있는데, 플레이오프 진출 그룹 (Playoffs = 1)의 평균 3점 성공률에 대한 구간 값이 그렇지 않은 그룹 (Playoffs = 0)에 비해 전반적으로 높고, 해당 구간에서 밀도 또한 높게 나타나는 것을 확인할 수 있으며, 3PA(3점 시도 횟수)에 대한 결과도 비슷한 양상을 나타냈다.

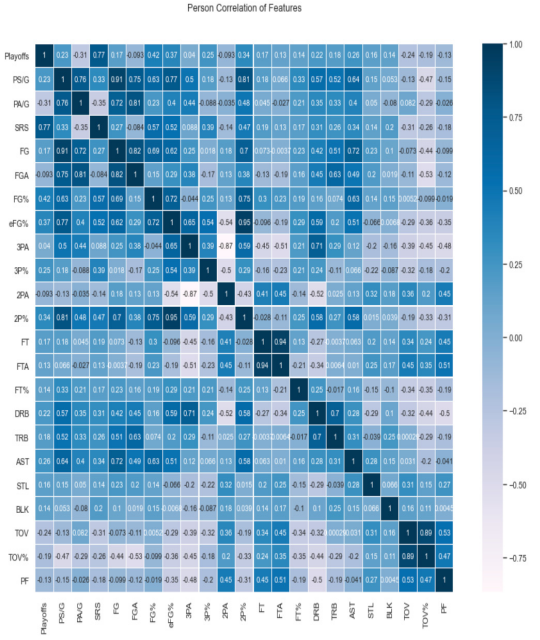


〈Figure 5〉 Comparison of 3-point Success rate by Groups Advancing and Dropping out of Playoffs

4.2. t 검정 및 분산분석

본 연구는 NBA의 플레이오프 진출에 영향을 미치는 변수들을 확인하고, 특히 3P 및 TOV와 관련된 속성을 중심으로 모델링 결과와 비교 분석하여 이러한 가정이 타당한가를 확인함에 그 목적이 있다.

모델 적용에 앞서 각 변수 간 상관관계를 분석하고, 플레이오프 진출 집단과 탈락 집단 간에 관심 대상인 주요 속성에 대해 평균의 차이가 있는지 확인하였다. 다음의 <Figure 6>은 수치형 변수의 상관관계를 보여주는 히트맵이다. 상관관계 분석의 결과 'SRS', 'FG%', 'eFG%'의 변수가 각각 0.77, 0.42, 0.37로 종속변수인 플레이오프 진출 여부와 가장 밀접한 양의 관계를 갖는 것으로 나타났다. 반면에 'PA/G' 및 'TOV' 변수는 각각 -0.31, -0.24로 음의 상관관계를 보였다.



〈Figure 6〉 Person Correlation of Features

플레이오프 진출 집단과 그렇지 않은 집단 간 주요 변수를 중심으로 평균에 차이에 관한 t 검정 결과는 다음 <Table 3>과 같다.

〈Table 3〉 Result of t-test

Variables	Levene test		t-test	
	Statistic	P-value	Statistic	P-value
eFG%	0.712	0.399	12.306	2.007e-32
3P%	0.0002	0.988	8.027	2.887e-15
TOV	1.604	0.206	7.612	6.420e-14
TOV%	1.019	0.313	6.089	1.635e-09

플레이오프 진출 집단과 그렇지 않은 집단 간 주요 변수를 중심으로 평균 차이를 검정하기 위해 데이터의 정규성 및 등분산성 검정(Levene test)을 수행하였고, 유의수준(p<.005)에서 모든 변수가 조건을 만족하였다. t 검정 결과 동일한 유의수준에서 통계적으로 유의한 것으로 나타났고, 이는 관심 대상 변수인 '3P'와 'TOV'에 있어서 플레이오프 진출과 미진출 집단 간 평균에 차이를 보이고 있으므로 모델링 결과를 통해 도출되는 주요 변수와 비교 분석하여 실제로 예측변수에 영향을 미치는 것인지 그 의미를 확인할 수 있다.

본 연구에서 설정한 종속변수는 플레이오프 진출 여부에 대한 속성임에도 불구하고, 실제 플레이오프 진출 및 승률과 같은 성적 지표와 관심 변수 간에 긴밀한 관계가 있는지 추가적인 검토를 위해 순위 정보를 바탕으로 생성한 변수인 Lev_Rk(시즌별 순위 수준)를 상, 중, 하의 3개 집단으로 구분하여 분산분석을 수행하였다.

분산분석 결과는 <Table 4>와 같으며, t 검정과 마찬가지로 유의수준(p<.005)에서 정규성, 등분산성의 조건을 만족하였으며, 순위로 분류한

상위, 중위, 하위 “세 집단의 평균이 모두 같다” 라는 귀무가설을 기각하고 “적어도 하나의 집단 간 해당 속성값 사이에 평균의 차이가 있다”라는 결론을 도출할 수 있었다.

〈Table 4〉 Result of ANOVA

Variables	Levene test		ANOVA	
	Statistic	P-value	F	P-value
eFG%	1.704	0.183	125.98	0.000
3P%	0.569	0.566	64.47	0.000
TOV	1.968	0.140	47.96	0.000
TOV%	1.466	0.231	29.60	0.000

4.3. 머신러닝 적용 결과 및 성능 평가

분석에 활용할 최종 입력 데이터 세트는 인스턴스 963건에 32개 컬럼으로 구성되며, 여기서 올해 시즌(2021-2022)에 해당하는 팀별 통계 데이터 30건은 테스트 세트(3.12%)로 제외하고 예측에 활용한다. 올해는 아직 NBA 시즌이 진행중이므로 수집된 데이터의 일자 기준으로 순위에 따라서 플레이오프 진출을 가정한다.

테스트 세트의 사이즈가 비교적 작고, 모델 훈련 시 발생할 수 있는 과적합(Overfitting) 방지를 위해 훈련 세트와 검증 세트를 나누어 5-fold 교차검증하였다. 5겹 교차검증 결과의 평균 모델 성능은 <Table 5>와 같다.

〈Table 5〉 Result of Cross-validation

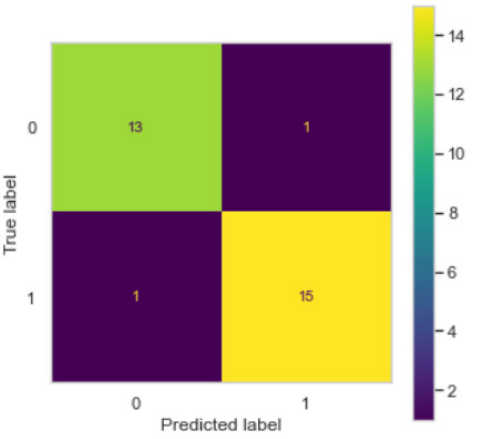
Models	precision_macro	recall_macro	f1_macro
Logistic Regression	0.88	0.88	0.88
Random Forests	0.88	0.88	0.88
AdaBoost	0.88	0.87	0.87
XGBoost	0.83	0.83	0.83

4가지 모델의 교차검증 결과 정밀도(Precision)와 재현율(Recall) 매크로 평균과 두 가지 지표를 모두 반영한 f1 매크로의 모든 지표에서 로지스틱 회귀와 랜덤 포레스트 모델이 가장 우수한 성능을 나타냈다. 모델 간에 약간의 성능 차이가 존재했지만 대체로 80% 이상의 높은 예측 결과를 보였다. 이러한 학습 결과를 토대로 이번 시즌의 입력 데이터를 실험 데이터로 하여 분석한 결과는 <Table 6>과 같고, 모든 모델에서 교차검증 결과보다 높은 성능을 보였으므로 훈련 세트에 과적합되지 않았음을 확인하였다.

테스트 데이터의 분석 결과에서 가장 우수한 성능으로 나타난 랜덤 포레스트 모델의 혼동 행렬(Confusion matrix)은 다음 <Figure 7>과 같다.

〈Table 6〉 Result of Analyzing test set

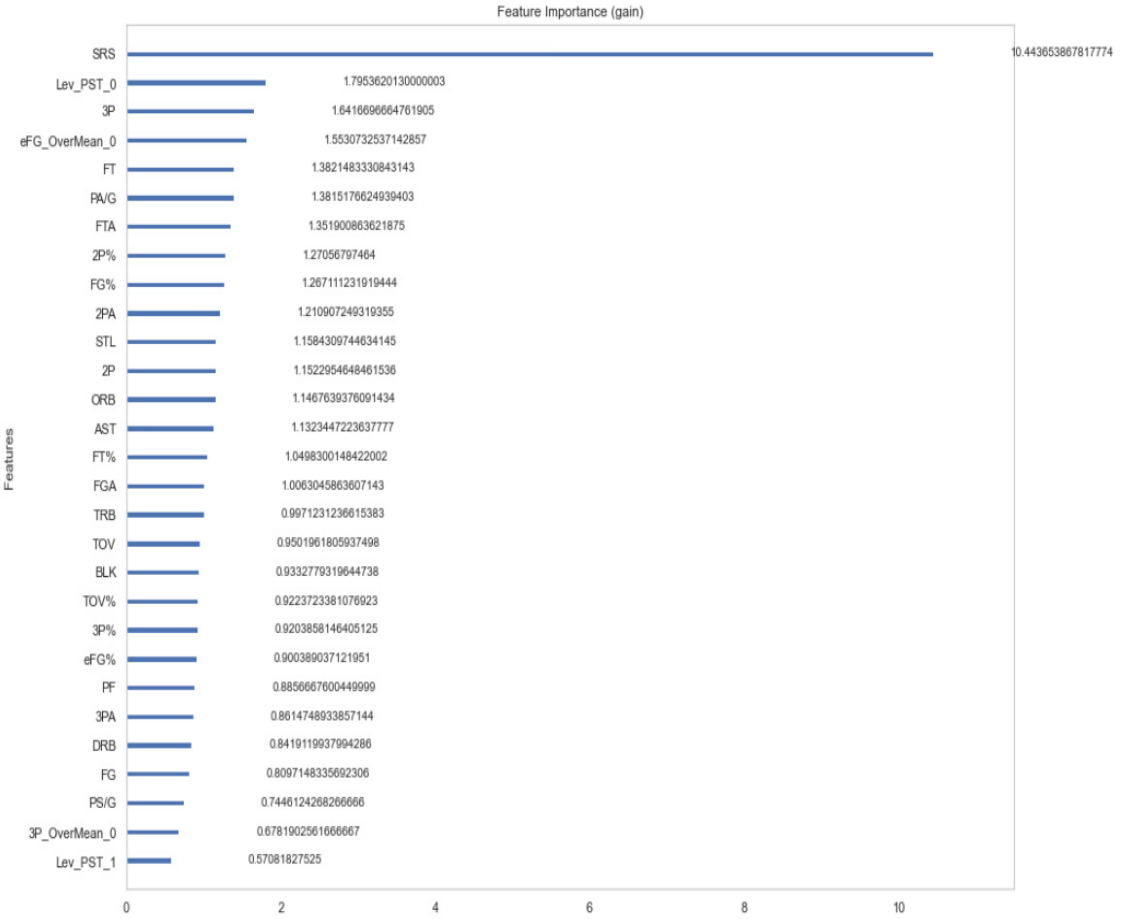
Models	precision_macro	recall_macro	f1_macro
Logistic Regression	0.9	0.9	0.9
Random Forests	0.93	0.93	0.93
AdaBoost	0.9	0.9	0.9
XGBoost	0.83	0.84	0.83



〈Figure 7〉 Confusion Matrix of RF

분석 결과를 요약하자면, 원시 데이터 자체가 30개의 팀별로 한 시즌에 80경기 진행에 대한 약 2400개의 경기기록(총 32시즌 데이터)을 요약하여 팀 통계로 제공하고 있어서 데이터의 품질 수준이 높고, 통계적 가정을 대부분 만족하고 있다. 따라서 적은 수준의 데이터 세트임에도 불구하고 모든 모델에서 대체로 높은 성능을 보이며, 특히 로지스틱 회귀와 같은 선형 모형에서도 비교적 높은 성능이 나타나는 것으로 보인다. 수치형 데이터 분석에서 비교적 우수한 성능을 나타내는 XGboost 모델의 성능이 상대적으로 낮은 이유로는 데이터 샘플에 비해 많은 변수가 투입되었고, 모형 복잡도에 적합한 튜닝을 통해 최적의 파라미터 세팅 없이 기본값으로 모델링한 결과로 볼 수 있다.

본 연구의 최종 목표는 머신러닝 기법을 이용하여 NBA의 플레이오프 진출에 영향을 미치는 주요 변수를 도출하고 관심 대상인 '3P'과 'TOV' 속성을 중심으로 비교 분석하는 것이다. 따라서 Shap Value 예측에 동일하게 사용된 결정



〈Figure 8〉 Feature Importance(gain) of XGboost

트리 기반의 XGboost의 특성 중요도(Feature importance)를 도출하고, 그 결과를 <Table 7>과 <Figure 8>에서 확인하였다.

최종적으로 분석을 통해 도출된 플레이오프 진출에 영향을 미치는 특성 중에 랜덤 포레스트 및 XGboost 각각에서 SRS가 가장 높은 중요도를 가지는 것으로 나타났으나, 본 연구의 관심 대상 변수인 3점 및 실책 속성과 관련된 'eFG_OverMean_1', '3P', 'eFG_OverMean_0', 'eFG%', '3P%', 'TOV' 등 다수의 변수가 중요한 특성으로 채택되었음을 확인할 수 있다.

〈Table 7〉 Summary of Feature Importance

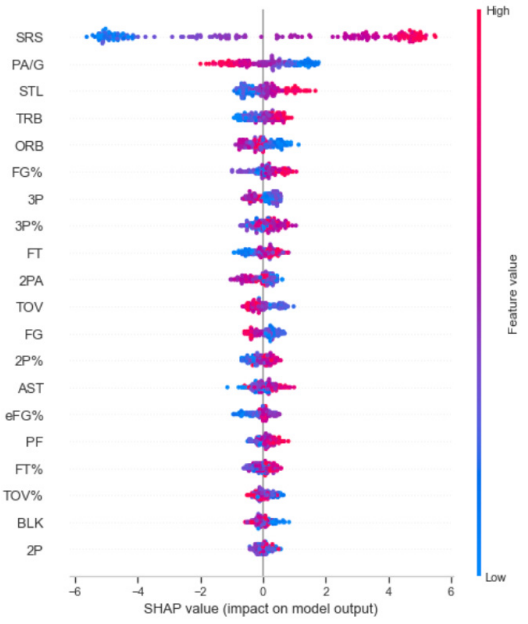
No.	Random Forest		XGboost	
1	SRS	0.344	SRS	0.255
2	PA/G	0.070	Lev_PST_0	0.044
3	eFG_Over Mean_1	0.055	3P	0.040
4	eFG_Over Mean_0	0.051	eFG_Over Mean_0	0.038
5	FG%	0.036	FT	0.034
6	2P%	0.034	PA/G	0.034
7	eFG%	0.030	FTA	0.033
8	STL	0.025	2P%	0.031
9	3P%	0.024	FG%	0.031
10	TOV	0.023	2PA	0.029

4.4. SHAP을 이용한 시각화

결정 트리 기반의 앙상블 모형에서 사용할 수 있는 특성 중요도의 분석 결과가 변수의 투입/제거에 따라 일관적이지 않고, 개별 속성이 어떠한 방향으로 영향을 미치는지 판단할 수 없다는 한계점을 갖는다. 이를 보완하기 위해 SHAP을 이용하여 결과에 대한 보다 상세한 설명이 가능하도록 시각화를 진행하였다. Shap value 예측에 XGboost를 사용하였고, 트리 기반의 Explainer를 생성하였다.

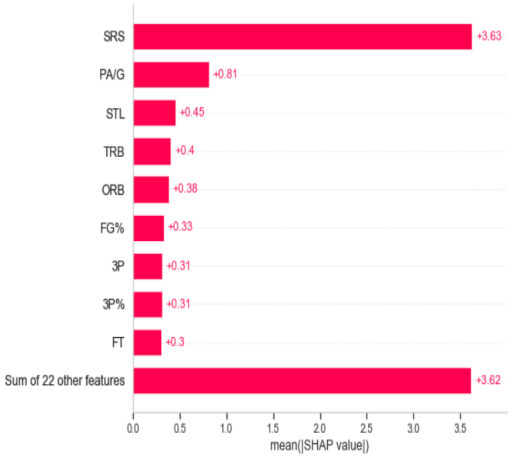
다음의 <Figure 9>는 Shap value에 대한 요약 플롯이며, Shap value란 모형 출력의 변화에 대한 특성의 영향 정도를 나타낸다. 플롯의 y축은 각 특성의 중요도 순으로 정렬되고, x축은 Shape value를 나타내며, 우측의 색 구분 막대는 특성값에 대한 표시로 붉은색에 가까워질수록 높은 값을 의미한다. 플롯에서 특성은 예측에 미치는 영향력에 따라 정렬됨에 따라 플레이오프 진출에 영향을 미치는 중요도 순서로 상위 20개의 변수가 출력되었고, 앞서 확인했던 특성 중요도와 비교하여 SRS 속성의 순위에는 변동이 없었다. SRS(Simple Rating System)는 득실 마진과 각 팀별 스케줄 강도를 감안하여 보정한 수치로써 실제 SRS가 높은 팀은 순위가 높은 경우가 대부분이다. 따라서 해당 변수는 다른 변수에 비해 압도적으로 플레이오프 진출 여부에 큰 영향을 미치고 있다고 할 수 있다. 그 밖에 관심 변수인 3점과 실책 속성에 관련된 변수인 '3P', '3P%', 'TOV', 'eFG%', 'TOV%'가 중요도 상위 20개의 변수에 포함되었다. 특징적으로 3점 성공 횟수(3P)와 3점 성공률(3P%)은 비슷한 중요도를 보이지만 반대 방향으로 플레이오프 진출에 영향력을 갖는 것으로 나타났다.

앞선 특성 중요도 산출 결과를 통해서는 단순하게 '3P'가 '3P%' 보다 중요도가 높다는 점만을 확인할 수 있었지만, SHAP을 통해서는 우선순위 외에도 예측변수에 개별 인스턴스별로 어떠한 방향으로 어느 정도의 영향을 미치고 있는지 직관적으로 확인할 수 있다.

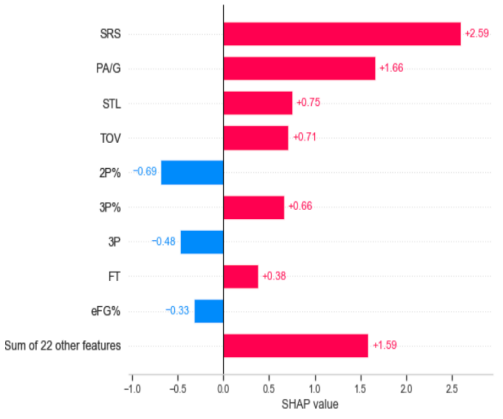


〈Figure 9〉 Summary Plot of SHAP

다음의 <Figure 10, 11>은 전체 데이터를 이용하여 각 특성의 영향력을 알 수 있는 전역적(Global) 방식과 개별 데이터에 대해 각 특성이 어떻게 영향을 미쳤는지 알 수 있는 지역적(Local) 방식에 대한 표준 막대 플롯이다. 전역적 방식은 모든 데이터에 대해 변수별 Shap value의 절대값 평균을 나타내고, 위의 요약 플롯과 유사한 의미를 내포한다. 반면 <Figure 11>은 전체 데이터 결과에서 특정 인스턴스에 각 특성이 어떻게 영향을 미치고 있는지 개별적인 확인이 가능하다. 지역적 출력 결과를 보면, 변수의 중요도 순위에서 실책 지표인 'TOV'가 상위권에 올랐고, 리바운드 지표인 'TRB', 'ORB'가 상위권에서 빠지며 중요도에 있어서 다른 양상을 나타내는 것을 확인할 수 있다.



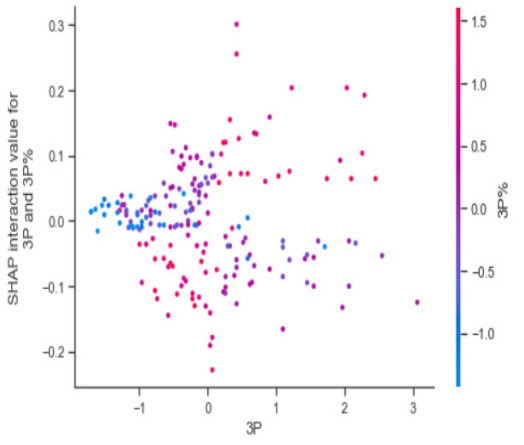
〈Figure 10〉 Bar Plot of SHAP(Global)



〈Figure 11〉 Bar Plot of SHAP(Local)

또한 개별 인스턴스에 따라 3점 지표로 분류하는 '3P', '3P%' 속성의 영향력과 그 방향성에서 차이가 발생하였는데, <Figure 12>의 의존도 플롯은 인스턴스별 두 속성의 주효과와 상호작용 효과를 나타내고 있다. 3점 성공률이 낮은 경우에 성공 횟수와 관계없이 두 변수 간 상호작용이 크지 않은 것으로 보이며, 3점 성공률이 높은 경우에는 3점 성공 횟수 구간에 따라 비교적 상

호작용 효과가 뚜렷하게 나타났다.



〈Figure 12〉 Dependence Plot of SHAP

5. 결론

본 연구는 미국 프로농구인 NBA의 플레이오프에 영향을 미치는 주요 변수를 도출함으로써 관심 변수인 3점과 실책 속성을 중심으로 비교 분석하였다. 머신러닝 분석에 앞서서 수집한 데이터에 대한 탐색적 분석을 통해 기존의 변수를 이용한 파생변수를 생성하였다. 3점 속성으로부터 유효 필드골 비율 (eFG%), 리그 평균 3점 성공률 (3P%_Mean), 리그 평균 3점 성공률 초과 여부 (3P_OverMean) 등의 변수를 생성하였고, 실책 속성으로부터 실책율 (TOV%) 변수를 생성하였다.

다음으로 입력 데이터에 대한 1차 정제 과정을 거쳐 무의미한 변수들을 제거하였고, 남은 변수에 대한 상관관계 분석, t 검정 및 분산분석을 수행하였다. ‘3점’ 및 ‘실책’ 속성과 관련된 요인들을 대상으로 두 집단인 플레이오프 진출/미진출 그룹 간 평균의 차이를 검정하고, 추가로 생성한 ‘Lev_Rk’ 변수를 통해 이러한 집단 간 평균 차이를 재확인하였다.

NBA 팀별 32개 시즌(1990/1991 ~ 2021/2022) 통계 데이터 중 올해 시즌의 데이터만을 테스트 세트로 활용하였고, 모델 훈련을 위해서는 훈련 세트와 검증 세트를 분할하여 5-fold 교차검증하였다. 교차검증 결과와 시험 세트를 이용한 최종 분석 결과를 비교하여 성능 지표에서 큰 차이가 없음을 확인함으로써 과적합 문제를 해결했다. 또한 로지스틱 회귀와 같은 선형 모형도 높은 성능이 나타나는 것으로 보였는데 이는 원시 데이터의 품질 수준이 높고, 통계적 가정을 만족하기 때문에 작은 수준의 데이터 세트임에도 불구하고 모든 모델에서 좋은 성능을 나타냈다.

본 연구는 단순히 머신러닝 기법을 이용하여 NBA의 경기 결과를 예측하거나 플레이오프 진출 여부만을 분류하는 것에서 그치지 않고, 입력 특성의 중요도를 파악하여 높은 중요도를 갖는 주요 변수에 본 연구의 관심 대상 변수가 포함되는지를 확인하였다. 실험 결과를 통해 3점과 실책 속성과 관련된 ‘eFG_OverMean_1’, ‘3P’, ‘eFG_OverMean_0’, ‘eFG%’, ‘TOV’ 등의 다수의 변수가 중요한 특성으로 채택된 것을 확인하였다.

최종적으로 Shap value의 시각화를 통해 특성 중요도의 결과만으로 해석할 수 없었던 속성별 영향력을 확인하였고, 변수의 진입/제거에 따른 중요도 산출에 일관성이 부족하다는 점을 보완할 수 있었다. SHAP의 시각화 결과로 도출된 주요 변수와 특성 중요도를 통해 도출된 변수가 완벽하게 같지는 않았지만, 본 연구에서 관심 대상으로 분류했던 3점 및 실책과 관련된 다수의 변

수가 NBA에서의 플레이오프 진출에 영향을 미치는 주요 변수로 포함되는 것을 확인하였다.

본 연구는 기존의 스포츠 데이터 분석 분야에서 다루었던 경기 결과, 플레이오프 및 우승 예측 등의 주제를 포함하고 분석을 위해 여러 머신러닝 모델을 비교 분석했다는 점에서 유사성이 있지만, 관심 속성을 사전에 설정하고, 이를 통계적으로 검증함으로써 머신러닝 분석 결과와 비교하였다는 측면에서 차이가 있다. 이어서 특성 중요도를 파악하기 위해 앙상블 모형을 적용함으로써 특성의 우선순위를 확인하고, 이를 통해 해석되지 않는 한계를 극복하기 위해 XAI 모델 중 하나인 SHAP를 이용하여 직관적인 결과를 제시함으로써 기존 연구와 차별화하였다. 3점 및 실책 특성과 관련된 관심 변수가 영향력 상위권에 다수 포함되었고, 각 특성별 해석을 통해 현대 농구에서 강조하는 빠르고 정확한 농구를 위한 3점과 실책 속성이 중요한 지표라고 간주할 수 있었다. 이러한 결과를 토대로 각 팀에서 중·장거리 슈팅 능력이 우수한 슈터에 대한 영입을 추진하거나 실책을 줄이기 위한 훈련 및 기록 중심의 성과 관리(인센티브/페널티 등)를 확대한다면, 그 실무적 활용도를 높일 수 있다.

분석에 활용한 원자료가 요약 통계를 반영하는 품질이 우수한 데이터이다 보니 모델에 대한 최적화 없이도 예측 성능이 대체로 높게 나왔다는 측면에서 한계점으로 남는다. 향후 위치, 센서, 이미지 등 다양한 데이터 소스를 활용한 실제 경기에서 발생하는 문제를 해결하기 위한 과제를 추후 연구로 남기며, 본 연구가 머신러닝을 이용하는 스포츠 분석 분야에서 다양한 접근을 위한 길라잡이가 되기를 기대한다.

참고문헌(References)

Albert, A. A., de Mingo Lopez, Luis Fernando, K. Allbright and N. Gomez Blas, "A Hybrid Machine Learning Model for Predicting USA NBA All-Stars," ELECTRONICS, Vol.11, No.1(2022), 97-112.

Wang, Y., W. Liu and X. Liu, "Explainable AI techniques with application to NBA gameplay prediction," Neurocomputing, Vol.483(2022), 59-71.

Araújo, D., M. Couceiro, L. Seifert, H. Sarmento and K. Davids, Artificial Intelligence in Sport Performance Analysis , Routledge, New York, 2021.

Bai, Z. and X. Bai, "Sports Big Data: Management, Analysis, Applications, and Challenges," COMPLEXITY, Vol.2021(2021), 6676297-6676307.

Chen, W., M. Jhou, T. Lee and C. Lu, "Hybrid Basketball Game Outcome Prediction Model by Integrating Data Mining Methods for the National Basketball Association," ENTROPY, Vol.23, No.4(2021), 477-490.

Geurkink, Y., J. Boone, S. Verstockt and J. G. Bourgois, "Machine Learning-Based Identification of the Strongest Predictive Variables of Winning and Losing in Belgian Professional Soccer," APPLIED SCIENCES-BASEL, Vol.11, No.5(2021), 2378-2388.

Jain, P. K., W. Quamer and R. Pamula, "Sports result prediction using data mining techniques in comparison with base line model," OPSEARCH, Vol.58, No.1(2021), 54-70.

Liu, S., Predicting NBA Playoffs Using Machine Learning, rScroll , 2021.

Wang, J. and Q. Fan, "Application of Machine

Learning on NBA Data Sets," Journal of physics. Conference series, Vol.1802, No.3(2021), 32036.

Yazbek, D., J. S. Sibindi and T. L. Van Zyl, "Deep Similarity Learning for Sports Team Ranking," 2021 SAUPEC/RobMech/PRASA (2021), 1-6.

Choi, Y. H. and K. H. Lee, "Analysis of Football Fans' Uniform Consumption: Before and After Son Heung-Min's Transfer to Tottenham Hotspur FC," J Intell Inform Syst, Vol.26, No.3(2020), 91-108.

Eom, H., J. Kim and S. Choi, "Machine learning-based corporate default risk prediction model verification and policy recommendation: Focusing on improvement through stacking ensemble mode," J Intell Inform Syst, Vol.26, No.2(2020), 105-129.

Geng, S. and T. Hu, "Sports Games Modeling and Prediction using Genetic Programming," In 2020 IEEE Congress on Evolutionary Computation (CEC)(2020), 1-6.

Han, D. Y., M. Hawkins and H. J. Choi, "Analysis of different types of turnovers between winning and losing performances in men's NCAA basketball," Journal of the Korea Society of Computer and Information, Vol.25, No.7(2020), 135-142.

Horvat, T. and J. Job, "The use of machine learning in sport outcome prediction: A review," Wiley interdisciplinary reviews. Data mining and knowledge discovery, Vol.10, No.5(2020), e1380.

Horvat, T., L. Havaš and D. Srpak, "The impact of selecting a validation method in machine learning on predicting basketball game outcomes," Symmetry, Vol.12, No.3(2020), art. no. 431.

Migliorati, M., "Detecting drivers of basketball successful games: an exploratory study with machine learning algorithms," Electronic Journal of Applied Statistical Analysis EJASA, Electron. J. App. Stat. Anal. Electronic Journal of Applied Statistical Analysis, Vol.13, No.2 (2020), 454-473.

Oh, J., Y. Lee and G. Kim, "Improvement of Solar Power Forecasting Using Interpretation of Artificial Intelligence," The transactions of The Korean Institute of Electrical Engineers, Vol.69, No.7(2020), 1112-1117.

Yi, J. H. and S. W. Lee, "Prediction of English Premier League Game Using an Ensemble Technique," KIPS Trans. Softw. and Data Eng., Vol.9, No.5(2020), 161-168.

Chen, Y., J. Dai and C. Zhang, "A neural network model of the NBA most valued player selection prediction," ACM International Conference Proceeding Series(2019), 16.

Horvat, T. and J. Job, "Importance of the training dataset length in basketball game outcome prediction by using naive classification machine learning methods," Elektroteh.Vestn. Electrotech.Rev., Vol.86, No.4(2019), 197.

Mandić, R., S. Jakovljević, F. Erčulj and E. Štrumbelj, "Trends in NBA and Euroleague basketball: Analysis and comparison of statistical data from 2000 to 2017," PLoS ONE, Vol.14, No.10(2019), 1-17.

Thabtah, F., L. Zhang and N. Abdelhamid, "NBA Game Result Prediction Using Feature Analysis and Machine Learning," Annals of Data Science, Vol.6, No.1(2019), 103-116.

Hsu, P. -., S. Galsanbadam, J. -. Yang and C. -. Yang, "Evaluating Machine Learning Varieties

for NBA Players' Winning Contribution," 2018 International Conference on System Science and Engineering (ICSSE)(2018), 1-6.

Lai, M., R. Meo, R. Schifanella and E. Sulis, "The role of the network of matches on predicting success in table tennis," J.Sports Sci., Vol.36, No.23(2018), 2691-2698.

Lam, M. W. Y., "ONE-MATCH-AHEAD FORECASTING IN TWO-TEAM SPORTS WITH STACKED BAYESIAN REGRESSIONS," JOURNAL OF ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING RESEARCH, Vol.8, No.3(2018), 159-171.

Rahman, M. H. A. A., A. Mustapha, N. Razali and R. Fauzi, "Bayesian approach to classification of football match outcome," International Journal of Integrated Engineering, Vol.10, No.6(2018), 155.

Bianchi, F., T. Facchinetti and P. Zuccolotto, "Role revolution: towards a new meaning of positions in basketball," Electronic Journal of Applied Statistical Analysis, Vol.10, No.3(2017), 712-734.

Giuliodori, P., "An artificial neural network-based prediction model for underdog teams in NBA matches," CEUR Workshop Proceedings, Vol.1971(2017), 73-82.

Kaur, H. and S. Jain, "Machine learning approaches to predict basketball game outcome," The 3rd International Conference on Advances in Computing Communication & Automation (ICACCA)(2017), 1-7.

Leicht, A. S., M. A. Gomez and C. T. Woods, "Team Performance Indicators Explain Outcome during Women's Basketball Matches at the Olympic Games," Sports (2075-4663), Vol.5, No.4(2017), 96-103.

Pai, P., L. ChangLiao and K. Lin, "Analyzing basketball games by a support vector machines with decision tree model," Neural Computing & Applications, Vol.28, No.12(2017), 4159-4167.

Cheng, G., Z. Zhang, M. N. Kyebambe and N. Kimbugwe, "Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle," Entropy, Vol.18, No.12(2016), 450-464.

Prasetio, D. and D. Harlili, "Predicting football match results with logistic regression," 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)(2016), 1-5.

Soto Valero, C., "Predicting win-loss outcomes in MLB regular season games-a comparative study using data mining methods," International Journal of Computer Science in Sport, Vol.15, No.2(2016), 91.

Al-Jarrah, O. Y., P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha, "Efficient Machine Learning for Big Data: A Review," Big Data Research, Vol.2, No.3(2015), 87-93.

Kempe, M., A. Grunz and D. Memmert, "Detecting tactical patterns in basketball: Comparison of merge self-organising maps and dynamic controlled neural networks," European Journal of Sport Science, Vol.15, No.4(2015), 249-255.

Lopez, M. J. and G. J. Matthews, "Building an NCAA men's basketball predictive model and quantifying its success," Journal of Quantitative Analysis in Sports, Vol.11, No.1(2015), 5-12.

Leung, C. K. and K. W. Joseph, "Sports Data Mining: Predicting Results for the College Football Games," Procedia Computer Science, Vol.35(2014), 710-719.

Zimmermann, A., S. Moorthy and Z. Shi, Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned, arXiv, 2013. Available at https://arxiv.org/pdf/1310.3607.pdf (Downloaded February 5, 2022).

Kim, S. H., J. W. Lee and M. S. Lee, "Estimating the determinants of victory and defeat through analyzing records of Korean pro-basketball," Journal of the Korean Data And Information Science Society, Vol.23, No.5(2012), 993-1003.

Schmidt, A., "Movement pattern recognition in basketball free-throw shooting," Human Movement Science, Vol.31, No.2(2012), 360-382.

Pak, S. I. and T. H. Oh, "The Application of Analysis of Variance (ANOVA)," Journal of Veterinary Clinics, Vol.27, No.1(2010), 71-78.

Kubatko, J., D. Oliver, K. Pelton and D. T. Rosenbaum, "A starting point for analyzing basketball statistics," Journal of Quantitative Analysis in Sports, Vol.3, No.3(2007), undefined.

Joseph, A., N. E. Fenton and M. Neil, "Predicting football results using Bayesian nets and other machine learning techniques," Knowledge-Based Syst., Vol.19, No.7(2006), 544-553.

Nunes, S., M. Sousa and d. E. Faculdade, Applying data mining techniques to football data from European championships, OpenAIRE, Europe, 2006.

Lee, G. B., "The factors of KBL team's playoff pass and winning percent," Korean Journal of Sport Science, Vol.15, No.3(2004), 41-50.

Abstract

Prediction of Key Variables Affecting NBA Playoffs Advancement: Focusing on 3 Points and Turnover Features

Sehwan An* · Youngmin Kim**

This study acquires NBA statistical information for a total of 32 years from 1990 to 2022 using web crawling, observes variables of interest through exploratory data analysis, and generates related derived variables. Unused variables were removed through a purification process on the input data, and correlation analysis, t-test, and ANOVA were performed on the remaining variables. For the variable of interest, the difference in the mean between the groups that advanced to the playoffs and did not advance to the playoffs was tested, and then to compensate for this, the average difference between the three groups (higher/middle/lower) based on ranking was reconfirmed. Of the input data, only this year's season data was used as a test set, and 5-fold cross-validation was performed by dividing the training set and the validation set for model training. The overfitting problem was solved by comparing the cross-validation result and the final analysis result using the test set to confirm that there was no difference in the performance matrix. Because the quality level of the raw data is high and the statistical assumptions are satisfied, most of the models showed good results despite the small data set. This study not only predicts NBA game results or classifies whether or not to advance to the playoffs using machine learning, but also examines whether the variables of interest are included in the major variables with high importance by understanding the importance of input attribute. Through the visualization of SHAP value, it was possible to overcome the limitation that could not be interpreted only with the result of feature importance, and to compensate for the lack of consistency in the importance calculation in the process of entering/removing variables. It was found that a number of variables related to three points and errors classified as subjects of interest in this study were included in the major variables affecting advancing to the playoffs in the NBA. Although this study is similar in that it includes topics such as match results, playoffs, and

* Graduate School of Technology & Innovation Management, Hanyang University
** Corresponding author: Youngmin Kim
Graduate School of Technology & Innovation Management, Hanyang University
222, Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea
Tel: +82-2-2220-2537, E-mail: yngmnkim@hanyang.ac.kr

championship predictions, which have been dealt with in the existing sports data analysis field, and comparatively analyzed several machine learning models for analysis, there is a difference in that the interest features are set in advance and statistically verified, so that it is compared with the machine learning analysis result. Also, it was differentiated from existing studies by presenting explanatory visualization results using SHAP, one of the XAI models.

Key Words : NBA Playoffs analysis, Machine learning, Statistical analysis, SHAP, Feature importance

Received : March 4, 2022 Revised : March 12, 2022 Accepted : March 22, 2022
Corresponding Author : Youngmin Kim

저 자 소 개

안 세 환
단국대학교에서 경영학 학사, 한양대학교 대학원에서 산업공학 석사 학위를 취득하였고, 현재 기술경영학 박사과정을 수료하였다.
한국생산기술연구원 스마트제조기술그룹에서 병역특례 연구원으로 근무한 바 있고, 주요 연구분야는 제조 데이터 분석, 통계적 품질관리, 머신러닝, 텍스트 마이닝 등이다.

김 영 민
한양대학교 산업공학과에서 학사, 석사 학위를 취득한 후 프랑스 Paris 6 대학 컴퓨터 공학과에서 석사, 박사 학위를 취득했다. Avignon 대학과 Lyon2 대학에서 박사후 연구원, 한국과학기술정보연구원에서 선임연구원으로 재직하였다.
2016년부터 한양대학교 기술경영전문대학원 교수로 재직하고 있다. 주요 연구분야는 기계학습, 확률 그래프모델, 정보 추출이다.