JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 4, pp. 11-18, April 2025
https://doi.org/10.9708/jksci.2025.30.04.011

A Comparative Study of Ensemble Learning Models for Predicting Attendance in the KBO League

Tai-Sung Hur*, Minsuk Oh**

*Professor, Dept. of Computer Science, Inha Technical College, Incheon, Korea
**Student, Dept. of Computer Science, Inha Technical College, Incheon, Korea

[Abstract]

This study developed and analyzed ensemble learning-based prediction models for forecasting attendance in the KBO League. Using KBO League data from 2022 to 2024, we collected variables such as team rankings, winning rates, consecutive wins/losses, search volume, stadiums, and home/away games, with the attendance ratio compared to stadium capacity set as the target variable. In the data preprocessing phase, Monday games were excluded, and the home/away attendance ratio was set to 7:3 to enhance model realism. Among various ensemble models compared, including Linear Regression, Random Forest, XGBoost, and LightGBM, the LightGBM model showed the best performance with an RMSE of 8.39 and R^2 Score of 0.783. Feature importance analysis revealed that online search volume (28.17%) and winning rate (25.17%) had the most significant impact on attendance, while team (10.57%) and day of the week (9.73%) also showed meaningful influence. Additionally, SHAP (SHapley Additive exPlanations) analysis provided insights into the directional impact of each variable on predictions, particularly revealing that the home/away factor had a stronger influence than expected through interactions with other variables. This study is significant in providing a practical prediction model that can assist KBO teams in establishing attendance strategies and making marketing decisions.

▸ **Key words**: KBO, Ensemble Learning, Attendance Prediction, LightGBM, SHAP analysis

[요 약]

본 연구는 KBO 리그의 관중 수요를 예측하기 위해 앙상블 학습 기반의 예측 모델을 개발하고 그 성능을 비교 분석하였다. 2022년 시즌부터 2024년 시즌까지의 KBO 리그 데이터를 활용하여 순위, 승률, 연승/연패, 검색량, 구장, 홈/원정 등의 변수를 수집하였으며, 홈 구장 대비 관중 비율을 목표 변수로 설정하였다. 데이터 전처리 과정에서 월요일 경기를 제외하고 홈/원정 경기의 관중 비율을 7:3으로 설정하여 모델의 현실성을 높였다. Linear Regression, Random Forest, XGBoost, LightGBM 등 다양한 앙상블 모델을 비교한 결과, LightGBM이 RMSE 8.39, R^2 Score 0.783로 가장 우수한 성능을 보였다. 특성 중요도 분석 결과, 온라인 검색량(28.17%)과 승률(25.17%)이 관중 동원에 가장 큰 영향을 미치는 것으로 나타났으며, 팀(10.57%)과 요일(9.73%)도 유의미한 영향을 미치는 것으로 확인되었다. 추가로 수행한 SHAP 분석을 통해 각 변수가 예측에 미치는 영향의 방향성을 파악할 수 있었는데, 특히 홈/원정 여부는 다른 변수들과의 상호작용을 통해 예상보다 더 큰 영향력을 미치는 것으로 나타났다. 본 연구는 KBO 구단의 관중 동원 전략 수립과 마케팅 의사결정에 실질적인 도움을 줄 수 있는 예측 모델을 제시하였다는 점에서 의의가 있다.

▸ **주제어**: KBO, 앙상블 학습, 관중 예측, LightGBM, SHAP 분석

• First Author: Tai-Sung Hur, Corresponding Author: Minsuk Oh
*Tai-Sung Hur (tshur@inhatc.ac.kr), Dept. of Computer Science, Inha Technical College
**Minsuk Oh (polpo444@naver.com), Dept. of Computer Science, Inha Technical College
• Received: 2025. 02. 25, Revised: 2025. 03. 31, Accepted: 2025. 04. 04.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

한국 프로야구(KBO 리그)는 국내 최고의 프로스포츠 리그로서 티켓판매를 통한 수익 창출과 함께 많은 팬들의 관심을 받고 있다. 한국 프로야구는 가장 먼저 시작된 프로스포츠로, 현재 대중적인 인기를 얻고 있다[1]. 특히 2024년 시즌은 흥행 열풍에 힘입어 국내 프로스포츠 사상 최초로 1000만 관중 시대를 열었다[2]. 프로스포츠에서 성공과 실패는 관람객 유인와 확보를 통한 이윤창출에 의해서 결정된다[3]. 또한 관중의 증가로 인하여 파생되는 수입원인 스폰서십, 구단 굿즈판매, 중계권료 등의 중요한 결정요인이 되고 있다[4].

스포츠 관중 분석에 관한 기존 연구들은 다양한 통계적 방법을 활용하여 관중 수요 예측을 시도해왔다. 권봉식은 관중 수요에 영향을 미치는 요인으로 라이벌 간의 경기, 경기 일정, 홈팀의 승률, 날씨 등을 제시하였다[5]. 김혁은 요일별 차이와 홈팀/원정팀 변수의 영향력을 정량적으로 분석하였으며[6], 조정환과 석부길은 Lasso 회귀, 랜덤 포레스트, XGBoost 등 머신러닝 기법을 활용하여 날짜, 날씨, 경기 상황 요인을 관측 변수로 이용한 관중 수요 예측을 실시하였다[7].

이러한 연구들이 의미있는 통찰을 제공했지만, 디지털 지표의 증가하는 중요성과 관중 동원에 영향을 미치는 다양한 요인들 간의 복잡한 상호작용을 고려하지 못했다. 특히 온라인 플랫폼과 소셜 미디어의 발달로 팬들의 야구 참여 방식이 변화하면서, 온라인 검색량이 실제 경기장 관중 수의 중요한 예측 지표가 될 수 있다는 점이 주목된다.

최근 머신러닝, 특히 앙상블 학습 방법의 발전은 관중 수 예측의 정확도를 높일 수 있는 가능성을 제시한다. 앙상블 학습은 여러 모델을 결합하여 단일 모델보다 더 신뢰성 있는 예측을 제공하며, 다수의 변수가 관련된 복잡한 예측 작업에 적합하다[8].

본 연구는 KBO리그의 경기 관중 수를 예측하기 위한 다양한 앙상블 학습 기반 모델을 개발하고 그 성능을 비교하는 것을 목적으로 한다. 기존의 요인 (팀 순위, 승률, 연승/연패 등)과 디지털 지표 (온라인 검색량)를 모두 고려하여 보다 포괄적인 예측 모델을 구축하고자 한다. 또한 홈/원정 경기의 관중 비율을 차별화하는 새로운 접근 방식을 도입하여 기존 연구의 한계를 보완하고자 한다.

II. Literature Review

2.1 Research Trends in Sports Attendance Prediction

관중 수 예측에 관한 연구는 전통적인 시계열 기법에서 최근의 기계학습 기반 접근 방식까지 점차 확장되어 왔다. 국내에서는 김형돈과 채진석이 시계열 모형을 활용하여 구단별 관중 수를 예측하였다[9]. 최근에는 딥러닝 및 머신러닝 기반의 접근이 활발해지고 있으며, 이수강 등은 네이버 검색량 데이터를 활용한 키워드 기반 관중 예측 모델을 제안하였고[10], 이다인 등은 NeuralProphet 시계열 모델을 통해 KBO 구단별 관중 수를 중장기적으로 예측하였다[11].

기존 연구들은 각기 다른 방법론을 통해 관중 수 예측의 가능성을 보여주었지만, 디지털 지표나 변수 해석력, 실시간성 등에서 한계가 존재하였다.

2.2 Predictive Models and Explainability Tools

본 연구에서 사용한 앙상블 모델(Random Forest, XGBoost, LightGBM)은 각각의 결정 트리를 통합하여 예측 성능을 높이는 기법으로, 비선형성과 변수 간 상호작용을 포착하는 데 효과적이다. 특히 LightGBM은 학습 속도가 빠르고 범주형 변수 처리에 강점을 가진다.

모델 해석 측면에서는 SHAP가 주목받고 있다. SHAP는 게임 이론의 샤플리 값을 기반으로 하여 개별 예측에 기여한 변수의 영향력을 정량적으로 분석할 수 있어, 블랙박스 모델의 투명성을 확보하는 데 효과적이다.

2.3 Contribution and Distinctiveness of This Study

기존 연구들은 주로 경기력, 경기 일정, 날씨 등 전통적 변수에 집중하거나, 단일 예측 모델 중심으로 접근한 경우가 많았다. 반면 본 연구는 온라인 검색량이라는 디지털 행태 지표를 정량 변수로 포함하고, 복수의 앙상블 모델을 비교하여 최적의 예측 구조를 설계하였다. 또한 SHAP 분석을 통해 변수별 영향력을 직관적으로 시각화함으로써 실무 적용 가능성을 높였다.

이처럼 기존 연구들은 다양한 방법론을 통해 관중 수 예측의 가능성을 보여주었으나, 디지털 지표의 활용, 변수 해석력, 실시간성 측면에서는 여전히 한계가 존재한다. 특히 NeuralProphet 기반 시계열 분석을 수행한 이다인 등의 연구[11]는 관중 수의 월별 추세와 시즌성(seasonality)을 강조하였으나, 변수 해석력이나 실시간 대응 측면에서는 아쉬움이 있었다. 이에 본 연구는 예측과 해석을 통합

하여 스포츠 산업 실무자에게 보다 유용한 의사결정 도구를 제공하고자 한다.

III. Methodology

1. Data Collection and Preprocessing

본 연구는 2022년부터 2024년 KBO 포스트 정규시즌 모든 경기를 대상으로 진행했다. 수집한 데이터는 크게 성적 데이터, 검색량 데이터, 관중 데이터로 구분된다.

팀별 순위, 승률, 연승/연패 등은 KBO 공식 홈페이지에서 Selenium으로 크롤링하여 수집하였다. 각 구단의 일자별 온라인 검색량 데이터는 데이터마케팅코리아에서 제공하는 데이터를 활용하였는데, MLB 리그 데이터는 제외하고 KBO 정규시즌 기간 동안 경기가 있었던 모든 일자들의 데이터만을 추출하여 CSV 파일로 저장 후 병합하였다.

관중 수 및 홈/원정 정보는 KBO 홈페이지에서 수집하고, 구장별 수용 인원은 위키피디아를 참조하여 관중 비율 변수로 활용하였다. 최종적으로 팀명과 일자를 기준으로 일자별 성적 데이터와 온라인 검색량 데이터를 병합하여 분석용 데이터셋을 구축하였다.

Table 1. Data Collection Results

Category	Data
Performance	Date, Rank, Team, Win Rate, Streak
Search Volume	Date, Team, PC/Mobile Search Value
Attendance	Date, Day, Home/Away, Stadium, Attendance

최종 구축된 데이터셋은 총 4,895개의 행과 11개의 열로 구성되었다. 변수들의 세부 정보는 Table 1과 같다. 수집된 데이터 중 구장(stadium), 홈/원정(home_or_away) 정보는 4,296개, 관중 수(attendance)는 2,149개, 관중 비율(att_ratio)은 2,144개의 데이터가 수집되었다. 관중 수와 비율 간 일부 누락은 2024년 청주구장 5경기에서 발생하였다. 관중 수와 관중 비율 데이터는 홈 경기에만 존재하기 때문에 전체 데이터의 약 44%를 차지한다. 관중 비율의 경우, 각 구장의 수용인원을 넘는 관중수 데이터로 인해 100이 넘는 값들은 100로 대체하였다. 검색량 데이터는 PC 및 모바일 검색량을 더한 합계를 이용한다.

Table 2. Dataset Variable Composition

Variable Name	Type	Example
date	string	2024.08.21
day	string	Tue,Wed...
rank	integer	1,2...
team	string	KIA,SSG..
win_rate	float	0.613
streak	string	2W,1L...
search_value	integer	122637
stadium	string	Jamsil
home_away	string	home
attendance	float	16000
att_ratio	float	100.0

월요일 경기의 경우, KBO리그에서 정규적으로 경기를 편성하지 않고 우천 취소된 경기의 보충경기나 특별한 상황에서만 진행되는 특성이 있어 데이터의 일관성을 위해 분석에서 제외하였다. 이로 인해 총 4,895개의 데이터 중 4,847개의 데이터가 분석에 사용되었다.

연승/연패(streak) 변수는 문자열 형태 (예:'2승','1패')로 되어있어 모델 학습을 위해 숫자형으로 변환하는 전처리 과정을 거쳤다. 또한, 요일(day), 팀(team), 구장(stadium) 등의 범주형 변수들은 모델 학습을 위해 레이블 인코딩(Label Encoding)을 적용하여 고유한 정수값으로 인코딩하였다.

본 연구의 데이터 처리 및 모델 학습은 Google Colab 환경에서 Pandas, Scikit-learn, XGBoost, LightGBM, SHAP 등의 라이브러리를 활용하여 수행되었다. 데이터 전처리와 통계 분석, 시각화는 Pandas, Seaborn, Matplotlib을 사용하였으며, 모델의 성능 평가 및 해석은 Scikit-learn의 내장 함수와 SHAP 라이브러리를 통해 진행하였다.

2. Exploratory Data Analysis

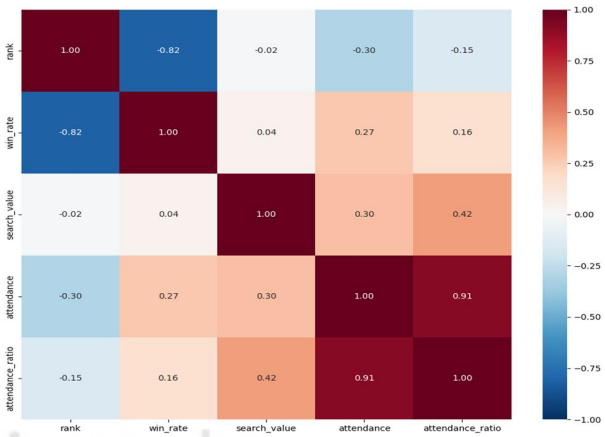

Fig. 1. HeatMap

Fig.1의 상관관계 분석 결과, 검색량은 관중 수 및 비율과 모두 유의미한 양의 상관관계를 보였으며, 순위는 승률과 강한 음의 관계를 나타냈다. 이는 온라인 관심도가 실제 관중 동원과 밀접히 연결되어 있음을 시사한다.

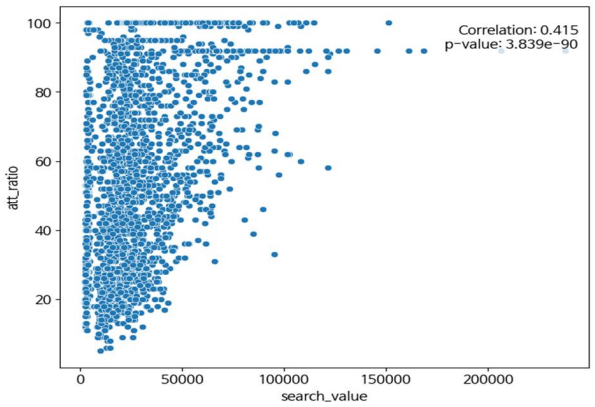

Fig. 2. Correlation between
Search value And Attendance Ratio

팀별 관중 동원력을 분석한 결과, 평균 관중 비율은 LG(64.8%)와 SSG(64.2%)가 가장 높았으며, NC(42.8%)가 가장 낮은 것으로 나타났다. 주목할 만한 점은 관중 비율 상위 2개 팀(LG, SSG)의 경우 관중 비율의 표준편차(각각 22.92, 21.99)가 다른 팀들에 비해 상대적으로 낮아, 안정적인 관중 동원력을 보여주었다.

검색량이 높은 팀(한화, 롯데, KIA)은 관중 비율도 높은 경향을 보여 상관관계를 뒷받침한다.

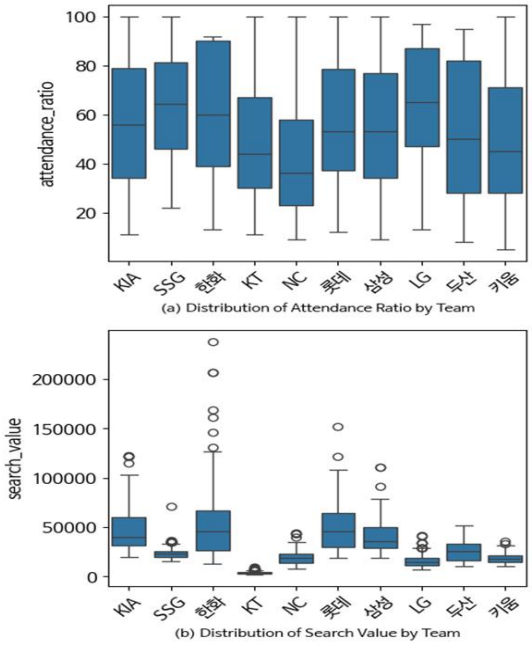

Fig. 3. Distribution of audience ratio and online search volume by KBO team

기초 분석 결과, 검색량과 관중 비율 간의 상관관계가 뚜렷하게 나타나 디지털 지표의 활용 가능성을 시사한다. 이러한 예비 분석 결과를 바탕으로, 다음 장에서는 보다 정교한 예측 모델 개발을 위한 방법론을 제시하고자 한다. 특히 기존 연구들에서 다루지 않았던 홈/원정 경기의 차별적 특성을 고려한 새로운 접근 방식을 통해 예측 모델의 정확도를 향상시키고자 한다.

IV. The Proposed Scheme

1. Basic Model: Attendance Ratio Prediction

본 연구는 KBO 리그의 관중 수요 예측을 위해 단계적인 모델 개발 접근 방식을 채택하였다. 첫 번째 단계로, 홈경기의 관중 비율만을 고려한 기본 모델을 개발하였다. 이를 위해 선형 회귀부터 앙상블 기법까지 다양한 머신러닝 알고리즘을 적용하고 그 성능을 비교 분석하였다. 이러한 결과는 Table 3에 요약되어 있다.

Table 3. Performance of Basic Models

Algorithm	RMSE	R^2 Score
Linear Regression	23.23	0.207
Decision Tree	22.48	0.257
XGBoost	17.45	0.552
Random Forest	16.99	0.576

가장 기본적인 접근 방식인 선형 회귀 모델은 RMSE (23.23), R^2 Score (0.207)의 성능을 보였다. 이는 관중 동원에 영향을 미치는 요인들 간의 관계가 단순한 선형성을 넘어선다는 것을 알 수 있다. Decision Tree 모델의 경우 RMSE (22.48) , R^2 Score (0.257)로 선형 회귀 모델보다는 나은 성능을 보였으나, 여전히 만족스러운 수준의 예측력 및 설명력을 보여주지 못했다.

앙상블 학습 기법을 적용한 결과, 예측 성능이 크게 향상 되었다. XGBoost는 성능이 우수하지만 학습 과정에서 상대적으로 높은 계산 비용이 발생할 수 있어, 데이터의 크기와 특성에 따라 성능이 다소 차이가 날 수 있다.[12]

Random Forest 모델은 RMSE (16.99), R^2 Score (0.576)으로 이 중에서 가장 우수한 성능을 보였다. 이는 Random Forest의 앙상블 학습 특성이 개별 의사결정 트리의 과적합을 효과적으로 제어하면서도, 전체적인 예측 성능을 향상시킨 결과로 해석된다. 특히, 변수의 무작위 샘플링 및 데이터 샘플링을 통해 데이터의 다양성을 확보하고, 변수 간 상호작용을 효과적으로 포착할 수 있었다.

또한 Out-of-Bag 검증을 통한 모델 평가에서 OOB점수 0.6으로 안정적인 성능을 보였다.

아래는 Random Forest 모델의 특성별 중요도를 시각화한 결과이다.

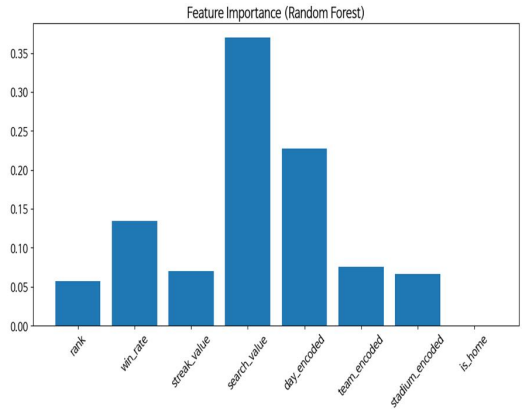

Fig. 4. Feature Importance

특성 중요도 분석 결과, 온라인 검색량 (search_value)이 0.35로 가장 큰 영향력을 보였다. 이는 디지털 시대에 팬들의 온라인 관심도가 실제 경기장 방문으로 이어지는 강한 연관성이 있음을 시사한다. 다음으로 요일(0.23)과 승률(0.14) 데이터가 예측 모델에 있어 비교적 큰 영향력이 있는 것을 확인 할 수 있었다.

그러나 이 기본 모델은 몇 가지 심각한 한계점을 가지고 있다. 첫째, 홈 경기 데이터만을 사용함으로써 전체 4,895개 데이터 중 약 44%인 2,144개의 데이터만을 활용하고 있다. 이는 모델의 학습에 사용할 수 있는 데이터의 양을 크게 제한하는 결과를 초래한다. 특히 원정 경기의 관중 동원 패턴을 완전히 배제함으로써, 실제 KBO 리그의 관중 동원 특성을 포괄적으로 반영하지 못하는 문제점이 있다.

둘째, 가장 우수한 성능을 보였던 Random Forest 모델의 R^2 Score가 0.576로 나타났다. 이는 모델이 데이터의 약 42.4%의 변동성만을 설명할 수 있다는 것을 의미한다. 또한, RMSE가 16.99로 측정되어, 최대 관중률이 100인 것을 감안한다면 예측값과 실제값 간의 차이가 큰 편이며, 이 결과는 모델 개선의 필요성을 시사한다.

이러한 한계점들을 극복하고 예측의 정확도를 더욱 높이기 위해, 다음 단계로 홈/원정 경기의 차별적 특성을 고려한 개선된 모델을 개발하였다. 개선 모델은 기존에 사용하지 않은 LightGBM 및 Gradient Boosting 등 다양한 알고리즘을 추가로 도입하여 모델의 성능을 더욱 향상시키고자 하였다. Gradient Boosting 알고리즘은 비선형 관계를 잘 처리할 수 있어, 기존 모델에서 간과될 수 있는 복잡한 패턴을 학습하는 데 유리하다. 특히 LightGBM은 대규모 데이터셋에서도 효율적인 학습이 가능하고 범주형 변수 처리에 강점이 있어, 홈/원정 경기의 특성을 반영한 새로운 모델에 적합할 것으로 판단하였다.

2. Improved Model Development

기본 모델의 한계를 극복하기 위해, 본 연구는 홈/원정 경기의 특성을 반영한 개선된 모델을 개발하였다. 이를 위해 먼저 홈/원정 관중 비율에 대한 차별적 접근을 시도하였다. KBO 리그의 특성상 홈 팀과 원정 팀의 관중 동원력에는 꽤나 큰 차이가 있다는 점을 고려하여, 원정 경기의 관중 비율을 해당 경기의 홈팀 관중 비율의 30%로 설정하였다. 이러한 7:3 비율의 설정은 KBO리그의 입장료 수익 배분 기준을 참고하였다. KBO에서 제공하는 관중 데이터는 전체 입장 관중수만 제공할 뿐 홈팀과 원정팀의 관중을 구분하여 집계하지 않고 있어 실제 홈/원정 관중의 정확한 비율을 파악하기 어렵다. 따라서 본 연구에서는 KBO 규정상 입장료 수익을 홈팀 72%, 원정팀 28%로 배분한다는 점에 근거하여 이와 유사한 7:3의 비율을 적용하였다.

2.1 Preprocessing of the Improved Dataset

본 연구의 데이터 전처리 과정은 다음과 같이 진행되었다. 먼저 월요일 경기 및 청주 구장에서 진행되었던 데이터는 전체 데이터에서 차지하는 비중이 매우 낮고 특수한 상황에서만 진행되는 경우가 많아서 분석 대상에서 제외하였다. 연승/연패 데이터의 경우 '3승', '2패'와 같이 문자열 형태로 되어 있어, 승리는 양수, 패배는 음수로 변환하여 수치화하였다. (예:'3승' -> 3, '2패' -> -2). 또한 홈/원정 여부는 이진 변수(binary variable)로 변환하여 홈경기는 1, 원정 경기는 0으로 더미 변수화하였다.

가장 중요한 전처리 과정으로 홈/원정 관중 비율 계산을 위해 새로운 변수를 생성하였다. 먼저 동일한 날짜의 동일한 구장에서 열린 경기를 찾아, 해당 경기의 홈팀과 원정팀을 구분하여 홈팀일 경우 기존 관중률 데이터에 0.7, 원정팀일 경우 0.3을 곱하여 추가하였다. 예를 들어, 특정 날짜에 잠실구장에서 두산(홈)과 LG(원정)의 경기가 열렸고, 이 경기의 실제 관중 비율이 80%였다고 한다면 두산(홈팀)은 56%, LG(원정팀)은 24%로 계산된다.

이러한 전처리 과정을 통해 사용 가능한 데이터의 수가 크게 증가하였다. 기존 모델에서 2,144개였던 데이터가 4,228개로 증가하여, 거의 두배에 가까운 데이터를 활용할 수 있게 되었다. 개선된 데이터셋을 바탕으로, 본 연구는 기존 알고리즘들에 LightGBM과 Gradient Boosting,

Ridge, Lasso, ElasticNet, AdaBoost 등을 추가하여 더 다양한 모델 비교를 시도하였다.

2.2 Algorithm Performance Analysis

본 연구에서 선택한 모델들은 예측 정확도, 해석 가능성, 계산 효율성을 기준으로 선정되었다. 선형 회귀는 기본 비교용으로 사용되었으며, Random Forest는 비선형성과 변수 상호작용을 잘 포착하고 과적합 방지에 강점을 지닌다. XGBoost와 LightGBM은 각각 정규화 및 범주형 변수 처리에 강점이 있어 대규모 예측에 적합하다.

추가적으로, 예측값과 실제값 간 오차를 직관적으로 평가하기 위해 MAE(Mean Absolute Error)를 도입하였다. 기존의 RMSE는 이상값에 민감하고 R^2 Score는 설명력 중심이기 때문에, MAE를 함께 고려함으로써 현실적인 예측 정확도를 보다 균형 있게 판단할 수 있도록 하였다.

Table 4. Performance of Improved Models

Algorithm	RMSE	R^2 Score	MAE
LightGBM	8.39	0.783	6.18
XGBoost	8.84	0.760	6.43
Random Forest	8.86	0.759	6.46
Gradient Boosting	9.08	0.746	6.81
AdaBoost	12.17	0.545	9.67
SVR	12.27	0.538	9.21
Linear Regression	12.79	0.497	9.98
Ridge	12.79	0.497	9.98
Lasso	13.02	0.479	10.17
ElasticNet	13.65	0.427	10.68

개선 모델들의 성능을 비교한 분석 결과, LightGBM이 RMSE (8.39), R^2 Score (0.783), MAE (6.18)로 가장 우수한 성능을 보였다. 이어서 XGBoost와 Random Forest가 비슷한 수준의 높은 성능을 보였다. 주목할만한 점은 앙상블 기반의 부스팅 계열 모델들 (LightGBM, XGBoost, Gradient Boosting)과 배깅 계열 모델 (Random Forest)이 모두 상위권의 성능을 보였다는 것이다.

반면, 선형 회귀 계열의 모델들 (Linear Regression, Ridge, Lasso, ElasticNet)은 상대적으로 낮은 성능을 보였다. MAE를 기준으로 볼 때, LGBM과 XGBoost가 가장 낮은 오차를 기록하였으며, Gradient Boosting도 유사한 수준의 정확도를 보여줬다. 반면, 선형 회귀 계열의 모델들과 AdaBoost는 MAE가 9 이상으로, 예측값과 실제값의 차이가 상대적으로 크게 나타났다. 특히 ElasticNet은 R^2 Score (0.427)로 가장 낮은 성능을 보였는데, 이는 L1, L2 정규화를 모두 적용하는 과정에서 모델의 표현력이 지나치게 제한되었을 가능성을 시사한다.

2.3 Interpretation of Feature Importance Analysis

여러 알고리즘을 적용한 모델 중 성능이 가장 우수했던 LightGBM 모델의 특성별 영향력을 분석한 결과는 Table 5와 같다. 온라인 검색량 (search_value)이 28.17%로 가장 높은 영향력을 보였으며, 승률 (win_rate)이 25.17%로 그 뒤를 이었다. 이 두 변수의 영향력 합이 53.34%로, 전체 변동의 절반 이상을 설명하는 것으로 나타났다. 이는 팬들의 온라인 관심도와 팀의 경기력이 관중 동원의 핵심 요인임을 시사한다.

중간 수준의 영향력을 보인 변수들로는 팀 구분(10.57%), 요일(9.73%), 연승/연패(8.60%)가 있다. 특히 팀 구분과 요일의 높은 중요도는 구단별로 차별화된 관중 동원 패턴이 있으며, 경기 일정이 관중 동원에 어느정도 영향을 미친다는 것을 보여준다.

구장(7.93%)과 순위(6.73%)는 상대적으로 낮은 영향력을 보였고, 홈/원정 여부(is_home)가 3.10%로 가장 낮은 영향력을 확인할 수 있었다. 이는 이미 홈/원정에 따른 관중 비율을 7:3으로 조정한 전처리 과정에서 해당 특성의 영향이 상당 부분 반영되었기 때문으로 해석된다.

Table 5. Feature Importance of LightGBM Model

Feature	Importance (%)
search_value	28.17
win_rate	25.17
team_encoded	10.57
day_encoded	9.73
streak_value	8.60
stadium_encoded	7.93
rank	6.73
is_home	3.10

2.4 Interpretation of SHAP Analysis

더 심층적인 특성 영향력 분석을 위해 SHAP 분석을 실시하였다. SHAP는 예측에 대한 개별 변수의 영향력과 방향성을 정량적으로 평가할 수 있는 도구이다. SHAP 값은 각 특성이 예측값을 증가시키는지(양의 값) 또는 감소시키는지(음의 값)를 나타내며, 그 크기는 영향력의 정도를 의미한다. Fig. 5은 각 특성의 SHAP 값 분포를 보여주는데, 점의 색상은 해당 특성의 원래 값의 크기를 나타내며, x축의 값은 예측에 대한 영향력의 크기와 방향을 나타낸다.

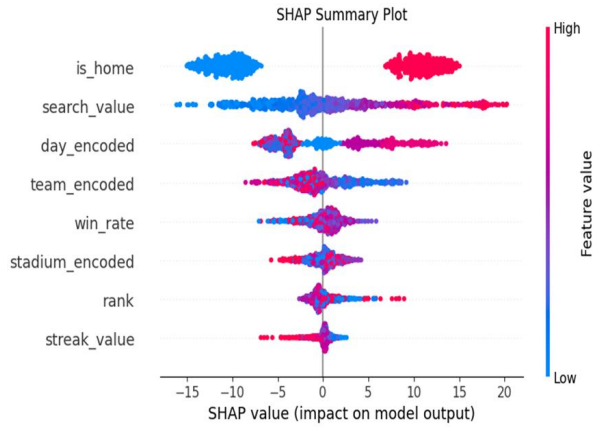

Fig. 5. SHAP Summary Plot

SHAP 분석 결과는 일반적인 특성 중요도와는 다소 다른 패턴을 보여주었다. 홈/원정 여부가 39.09%로 가장 큰 영향력을 보였으며, 검색량(20.02%)과 요일(16.98%)이 그 뒤를 이었다. 특히 주목할 만한 점은 변수들의 영향 방향성이다. 홈/원정 여부, 검색량, 요일은 관중 동원에 대체로 양의 영향을 미치는 것으로 나타났다. 반면 팀 구분(-0.36), 순위(-0.02), 연승/연패(-0.06)는 평균적으로 음의 영향을 미치는 것으로 분석되었다.

홈/원정 여부(is_home)는 다른 특성들과 달리 -15와 15 부근에 집중적으로 분포되어 있는데, 이는 이진 변수의 특성상 관중 동원에 명확한 양(홈)과 음(원정)의 영향을 미치는 것을 보여준다. 검색량은 -15에서 20까지 넓은 범위의 영향력을 보였는데, 이는 온라인 관심도가 상황에 따라 관중 동원에 매우 다른 영향을 미칠 수 있음을 시사한다. 요일과 팀은 비교적 고른 분포를 보이고 있지만, 특정 요일과 특정 팀의 경우 더 큰 영향력을 미치는 것으로 나타났다. 승률(win_rate)의 경우 높은 값(빨간색 점)이 대체로 양의 SHAP 값을 가져, 높은 승률이 관중 동원에 긍정적인 영향을 미치는 것을 확인할 수 있다.

이러한 SHAP 분석 결과는 관중 동원에 영향을 미치는 요인들의 복잡한 상호작용을 보여준다. 특히 일반적인 특성 중요도에서는 상대적으로 낮은 영향력을 보였던 홈/원정 여부가 SHAP 분석에서는 가장 큰 영향력을 보인 것은, 이 변수가 다른 변수들과의 상호작용을 통해 관중 동원에 더 큰 영향을 미칠 수 있음을 의미한다.

3. Ensemble Model Evaluation

본 연구에서 비교한 여러 알고리즘 중 앙상블 기반의 모델들, 특히 LightGBM이 가장 우수한 성능을 보였다. 선형회귀 모델이 R^2 Score 0.497을 기록한 것에 비해, LightGBM은 0.783으로 현저히 높은 설명력을 보였다. 이는 관중 동원에 영향을 미치는 요인들 간의 관계가 단순한 선형성을 넘어서는 복잡한 패턴을 가지고 있어, 비선형적 관계를 앙상블 모델이 효과적으로 포착했기 때문으로 해석된다.

앙상블 학습은 여러 학습 알고리즘을 조합해 더 강력한 예측 모델을 만드는 방법이다. 본 연구에서 사용된 LightGBM, Random Forest, XGBoost는 모두 결정 트리를 기반으로 하지만, 결합 방식이 다르다. Random Forest는 배깅(Bagging) 방식을 사용해 여러 트리의 예측을 평균내며, LightGBM과 XGBoost는 부스팅(Boosting) 방식을 통해 이전 모델의 오차를 보완한다.

LightGBM은 그래디언트 부스팅 방식의 앙상블 알고리즘으로, 리프 중심 트리 분할(leaf-wise tree growth) 방식을 채택해 더 빠른 학습이 가능하다. 또한 LightGBM은 카테고리형 변수에 대한 자체적인 인코딩을 지원하여 특성의 고유한 특성을 보존하면서 학습이 가능하다. 이는 본 연구의 데이터셋처럼 범주형 변수(팀, 구장, 요일 등)와 연속형 변수(승률, 검색량 등)가 혼재된 경우에 큰 장점이 된다. SHAP 분석 결과에서 확인했듯이, 홈/원정 여부와 같은 변수는 다른 특성들과의 상호작용을 통해 예상보다 더 큰 영향을 미칠 수 있는데, LightGBM은 이러한 특성 간의 비선형적 관계와 상호작용을 효과적으로 포착하여 더 정확한 예측을 가능하게 한다.

이러한 앙상블 학습의 장점은 본 연구의 결과에서도 명확히 드러난다. 모델의 예측 오차를 살펴보면, 평균 오차가 -0.04로 0에 매우 가깝고 오차의 표준편차가 8.39로 나타났다. 이는 모델이 전반적으로 편향되지 않은 예측을 수행하며, 대부분의 예측이 합리적인 범위 내에서 이루어지고 있음을 보여준다.

V. Conclusions

본 연구는 KBO 리그의 관중 수요를 예측하기 위해 앙상블 학습 기반 모델을 개발하고, 기존 통계적 요인뿐 아니라 온라인 검색량과 같은 디지털 지표를 결합하여 예측 정확도를 향상시켰다. 또한 홈/원정 경기의 차별적 특성을 반영함으로써 모델의 현실성을 높였다.

모델 비교 결과, LightGBM이 R^2 Score 0.783, RMSE 8.39, MAE 6.18로 가장 우수한 성능을 보였으며, 검색량(28.17%)과 승률(25.17%)이 주요 변수로 나타났다. SHAP 분석을 통해 각 변수의 영향 방향과 상호작용도 시각적으로 확인할 수 있었다.

본 연구의 실무적 기여는 다음과 같다. 예측 모델을 활용해 구단은 낮은 관중이 예상되는 경기일에 마케팅, 할인 이벤트, 인기 선수 기용 등의 전략을 사전 수립할 수 있다. 또한 관중 예측은 인력 배치, 물품 준비, 스폰서 단가 책정 등 운영 전반에 걸쳐 활용 가능하다. 이는 수익 극대화와 팬 경험 향상에 직접적인 도움을 줄 수 있다.

한편, 본 연구는 날씨, 티켓 가격, 이벤트 여부 등 외생변수들을 반영하지 못한 한계가 있다. 추후 연구에서는 외부 환경 변수, 소셜 미디어 반응, 실시간 검색량 변화 등을 반영한 다변량 예측 모델로 확장하여 정확도를 높이는 방향이 요구된다. 아울러 팀별 마케팅 전략 또는 경기 특성까지 반영한 맞춤형 관중 예측 연구로도 이어질 수 있다.

REFERENCES

[1] Sang-Hun Sung, "Analyses on Determinants of Attracting Spectators in Korean Professional Baseball League : A Study on Competitive Balance," 2019.

[2] Ha, Mu-rim, "KBO League Reaches Historic 10 Million Spectators for the First Time!," KBS News, September 15, 2024. https://news.kbs.co.kr/article/view.do?ncd=8060144

[3] Lim, Nam-Kyun, Jung, Mun-Yong, and Chung, Tae-Wook, "The Relationships among Organizational Trust of Professional Sports, Referee Trust, Spectator Satisfaction and Customer Citizenship Behavior," Journal of Sport and Leisure Studies, Vol. 69, pp. 155-164, 2017. DOI: 10.51979/KSSLS.2017.08.69.155

[4] Lee Chang-Sub, Kim Dae-Hee, and Hwamg Sung-Ha "Consumer Sentiment of Korean Professional Baseball Spectators in Terms of Big Data," The Korean Journal of Sport, Vol. 17, No. 2, pp. 881-889, 2019.

[5] Kwon Tae-Won, Park Seung-Hyun, and Kwon Bong-Sik, "Factors Attracting Attendance at Korean Professional Baseball Using Decision Tree Technique," Korean Journal of Sports Science, Vol. 15, No. 1, pp. 433-443, 2006.

[6] Kim Hyeuk, "Prediction of the number of attendances in the home team according to the visiting team and the day in Korean Baseball League," Korean Journal of Sport Management, Vol. 21, No. 6, pp. 85-96, 2016.

[7] Cho, Jung-Hwan, and Seok, Boo-Gil, "The Development prediction model of Korea Professional Baseball league spectator using machine learning," Korean Journal of Sports Science, Vol. 32, No. 5, pp. 547-558, 2023. DOI: 10.35159/kjss.2023.10.32.5.547

[8] Polikar, Robi, "Ensemble Learning," Ensemble Machine Learning: Methods and Applications, Springer, pp. 1-34, April 2012. DOI: 10.1007/978-1-4419-9326-7_1

[9] Hyung Don Kim, and Jin Seok Chae, "Prediction of the Number of Spectators for the Pro-baseball Club Using a Time Series Model," The Korean Journal of Measurement and Evaluation in Physical Education and Sports Science, Vol. 14, No. 3, pp. 57-68, 2012.

[10] Lee, Sugang, Jeon, Chaewon, Choi, Su-A, Kim, Seungbeom, and Jun, Hongbae, "A Prediction Model for the Number of Spectators in the Korean Professional Baseball League Using Keyword Search Volume," Korean Journal of Sport Management, vol. 29, no. 4, pp. 40-51, 2024.

[11] Lee, Da-In, Heo, Do-Hwi, and Chung, Ji-Young, "Prediction of Team-Specific Attendance in the KBO League Through Time Series Analysis," Korean Journal of Sports Science, Vol. 34, No. 1, pp. 161-176, 2025. DOI: 10.35159/kjss.2025.2.34.1.161

[12] Chen, Tianqi, and Guestrin, Carlos, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, San Francisco, USA, August 2016. DOI: 10.1145/2939672.2939785

Authors

Tai-Sung Hur received the B.S degree in Dept. of Computer Science from Inha University in 1984, and M.S degree in Dept. of Computer engineering from Soongsil University in 1987, and Ph. D. degree in Dept. of Computer engineering from Inha University in 1992. Dr. Hur has over 35 years of computer education. He is currently a Professor in the Dept. of Computer Science, Inha Technical College. He is interested in Data Science, Big data, Database and Internet of Things.

Minsuk Oh received B.S. degree in 2025 from the Department of Computer Science Inha Technical College, Incheon Korea. His research interests include Data Science, Big data and Data engineering.