The Journal of the Convergence on Culture Technology (JCCT)
Vol. 10, No. 3, pp.131-135, May 31, 2024. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2024.10.3.131
JCCT 2024-5-16

딥러닝 기반 일별 야구 관중 수 예측

Deep Learning-Based Daily Baseball Attendance Predcition

이현희*, 손서영**, 박민서***

Hyunhee Lee*, Seoyoung Sohn, Minseo Park*****

요 약 한국에서 야구는 프로 스포츠 종목 중 가장 많은 관중 수를 동원하고 있다. 특히 수입 대부분이 입장 수입이기 때문에 관중 수가 무엇보다 중요하다. 기존 연구는 타 종목이나 모든 구장을 동시에 고려하고 있어 구장 별 관중수를 예측이 쉽지 않다는 한계가 존재한다. 예를 들어 기아 타이거즈는 국내 구단 중 가장 높은 원정 수입을 보이는데에 반해 낮은 홈 수입을 보인다. 따라서, 본 연구에서는 딥러닝(Deep Learning)을 사용하여 기아 타이거즈의 광주 - 기아 챔피언스 필드의 일별 관중 수를 예측하고자 한다. 2018년~2023년의 광주 - 기아 챔피언스 필드의 일별 관중 수와 날짜, 날씨, 팀과 관련된 변수를 수집하고 전처리한다. 전처리 한 데이터를 활용하여 일별 관중 수를 예측하는 딥러닝기반 선형 회귀모델을 제안한다. 본 연구를 통해 구단의 수익 증대를 위한 기초 자료로 활용할 수 있을 것으로 기대한다.

주요어 : 딥러닝, 야구, 관중 수 예측

Abstract Baseball attracts the largest audience among professional sports in Korea. In particular, attendance is the primary source of income in baseball. Previous studies have limitations in reflecting the characteristics of individual stadium. For instance, the KIA Tigers exhibit the highest away game revenue among domestic teams, but they show lower home game earnings. Therefore, we aim to predict the daily attendance at the Gwangju-KIA Champions Field of the KIA Tigers using deep learning. We collected and preprocessed daily attendance, dates, weather, and team-related variables for Gwangju-KIA Champions Field from 2018 to 2023. We propose a deep learning-based linear regression model to predict the daily attendance. We expect that the proposed deep learning model will be used as basic information to increase the club's revenue.

Key words : Deep Learning, Baseball, Prediction of Daily Attendance

Ⅰ. 서 론

현대에 들어 국내 프로 스포츠 관람은 새로운 여가 활동으로 자리매김하였다[1]. 주 5일제 근무로 인해 여가 시간이 증가하고, 스포츠 관람에 대한 관심이 늘어나며 스포츠를 통해 스트레스를 풀고자 하는 국민들이 증가하였기 때문이다. 이로 인해 프로 스포츠 시장은 점차 성장하는 모습을 보여주었다. 특히, 1982년 가장

*준회원, 서울여자대학교 데이터사이언스학과 학부생(제1저자)
**준회원, 서울여자대학교 데이터사이언스학과 학부생(참여저자)
***정회원, 서울여자대학교 데이터사이언스학과 조교수(교신저자)
***Corresponding Author: mpark@swu.ac.kr
Dept. of Data Science, Seoul Women's Univ, Korea
Received: March 4, 2024 / Revised: April 10, 2024
Accepted: April 20, 2024
접수일: 2024년 3월 4일, 수정완료일: 2024년 4월 10일
게재확정일: 2024년 4월 20일

먼저 출범한 프로 야구는 출범 이후 연고지를 바탕으로 높은 시장 성장세를 보여주고 있다[2]. 2017시즌 약 840만명의 관중을 동원하며 프로 스포츠 역대 최다 누적 관중을 기록하였다[3]. 또한, 코로나 19 팬데믹 이후인 2023시즌에도 약 810만 명의 관중을 동원하며 역대 세 번째로 많은 누적 관중을 기록하였다[4]. 이를 통해 국내 프로 스포츠 시장에서 한국 프로 야구를 향한 국민들의 관심을 알 수 있다.

프로 스포츠의 관중 수는 인기를 나타내는 척도이자 구단의 경영과 직결되는 요소이다[5]. 특히, 프로 야구는 입장 수입이 전체 수입의 대부분을 차지하며, 부대시설 이용료, 중계권료, 광고 수입 등의 부가적인 수입의 경우 입장 수입으로 인해 파생된다[6]. 따라서, 관중 수의 증감이 구단의 수입으로 직결되는 프로 야구에서 높은 정확도의 관중 수 예측은 필수적이다.

그러나 기존 연구 대부분이 야구 외에 다른 스포츠 종목을 함께 고려하거나, 하나의 구장이 아닌 전 구장을 동시에 고려한다. 스포츠는 종목마다 경기 특성이 다르고, 야구장마다 수용 인원과 좌석 형태가 다르기 때문에 각 구장에 초점을 맞춘 일별 관중 수 예측 모델이 필요하다.

한국 프로 야구는 총 10개의 구단이 팀을 이루고 있다. 그 중, 기아 타이거즈는 2022년 KBO(Korea Baseball Organization) 리그 시청률 상위 10개 경기에 모두 포함될 만큼 많은 팬을 보유하고 있다. 그러나, 2022 정규 시즌 원정 경기 기준 입장 수입이 구단 중 가장 높은데 반해 홈경기 입장 수입은 5위에 미친다[4]. 이는 기아 타이거즈가 원정 경기로는 많은 수입을 얻지만 홈 구장인 광주 - 기아 챔피언스 필드에서는 큰 수입을 얻지 못하는 것을 의미한다.

따라서 본 연구는 10개 구단 중 우선 광주-기아 챔피언스 필드의 일별 관중 수를 예측하고자 한다. 선행연구를 기반으로 다양한 변수를 고려하여 딥러닝(Deep Learning) 기반 예측 모델을 제안한다. 딥러닝은 층을 깊게 쌓을 수 있어 다양한 변수를 고려할 수 있고, 주요 테이터 특징을 자동으로 추출할 수 있다는 장점이 있다[7, 8].

본 논문의 구성은 다음과 같다. 제 2장은 프로 스포츠 관중 수 예측과 관련된 선행 연구에 대해 기술하였다. 제 3장에서는 본 연구에서 제안하는 딥러닝 기반 일별 광주 - 기아 챔피언스 필드 모델을 설명한다. 제 4장에서는 결과 및 검증을 서술하고, 제 5장에서는 결론을 언급한다.

II. 선행 연구

프로야구 관중 수 예측 및 관중 수에 영향을 미치는 요인을 분석한 선행 연구들을 살펴보았다.

박진욱 외[5]는 프로야구 구장별 일일 관중 수를 예측하기 위해 독립 변수로써 2015년 3월부터 9월까지의 시간, 날씨, 지역, 팀별 특성, 누적 경기 성적 및 대중 요소 변수들을 사용하였다. 수집한 변수들을 활용하여 관중 수 예측 모델을 만들기 위해 다중 선형 회귀(Multiple Linear Regression)와 인공신경망(Artificial Neural Network, ANN) 모델을 활용해 모델링하였다. 인공신경망이 다중 회귀에 비해 평균적으로 26.3% 뛰어난 성능을 보임을 확인하였다.

김혁 외[9]는 프로야구, 프로농구, 남자 프로배구, 여자 프로배구 각각의 일일 관중 수를 예측하는 모델을 제안하였다. 2015년부터 2018년까지의 시간, 팀, 팀 성적 요소 변수들을 활용해 선형 회귀 모형, 랜덤 포레스트(Random Forest), XGBoost, 인공신경망 모델로 모델링하였다. 모델링 결과, 인공신경망의 성능이 가장 높은 성능을 보임을 확인하였다.

조정환 외[10]는 프로야구 구장별 일일 관중 수를 예측하기 위해 2017년부터 2019년까지의 날짜, 팀, 경기 상황, 날씨 요소 변수들을 사용하였다. 수집한 변수들을 활용하여 관중 수 예측 모델을 만들기 위해 선형 회귀, 랜덤 포레스트, XGBoost를 활용해 모델링하였다. 그 결과, XGBoost로 만든 모델의 성능이 가장 높게 나타남을 확인하였다.

프로야구 관중 수 예측에 관한 선행 연구는 대부분 다양한 종목 또는 프로야구의 모든 팀을 한 번에 분석하고 있어 구장별로 최적화된 모델을 만들 수 없다는 한계가 있다. 또한, 대부분의 연구가 머신러닝(Machine Learning)과 인공신경망을 기반으로 진행되었으며, 다양한 변수를 고려할 수 있는 딥러닝 모델을 활용한 연구는 부족한 실정이다. 따라서 본 연구는 광주 - 기아 챔피언스 필드의 일일 관중 수를 예측하는 딥러닝 기반 선형 회귀 모델을 제안한다.

III. 모델 설계

본 연구에서는 광주 - 기아 챔피언스 필드의 일별 관중 수를 예측하기 위한 딥러닝 기반 선형 회귀 모델을 설계하였다.

1. 데이터 수집

프로스포츠 정보 광장에서 2018년부터 2023년까지의 광주 - 기아 챔피언스 필드의 일별 관중 수 데이터를 수집하였다. 또한, 2018년부터 2023년까지의 광주 - 기아 챔피언스 필드의 일별 관중 수 예측을 위한 선행 연구를 바탕으로 13개의 독립 변수를 수집하였다[10, 11] 수집한 독립 변수는 날짜 관련 요소, 날씨 관련 요소, 팀 관련 요소로 구성된다. 날짜 요소에는 연도, 월, 요일, 공휴일, 시간대, 개막전이 속하며, 팀 요소에는 상대 팀, 상대 팀 순위, 홈 팀 순위가 포함된다. 날씨 요소에는 평균 기온, 평균 습도, 미세먼지, 평균 강수량이 속한다.

2. 데이터 전처리

수집한 2018년부터 2023년까지 데이터 중 코로나 19의 영향을 받은 2020년, 2021년은 제거하였다. 2020년과 2021년은 코로나 19로 인해 사회적 거리두기가 실시됨에 따라 관중 수의 제한을 두었기 때문이다. 두 개의 연도를 제거한 후, 우천으로 취소되어 관중 수가 0명인 경기의 관련 데이터를 제거하였다.

독립 변수 중 날짜 요소의 연도는 2018년, 2019년, 2022년, 2023년 총 4개의 범주, 월은 3월~10월 총 8개의 범주, 요일은 월요일~일요일의 총 7개의 범주로 구성하였다. 공휴일은 주말과 공휴일을 포함하여 공휴일인 경우 1, 아니면 0으로 변환하였다. 시간대는 오후 경기(5시 이후)일 경우 1, 아닐 경우는 0으로 표기하였다. 개막전은 개막전 시리즈라고 할 수 있는 시즌 첫 토요일~일요일 경기에 해당할 경우는 1, 그 외는 0으로 변수를 생성하였다. 팀 요소의 상대 팀은 홈 구단인 기아 타이거즈를 제외한 9개의 상대 팀을 범주형 변수로 생성하였고, 상대 팀 순위의 경우 각 시즌 별 해당일 이전까지 상대 팀의 순위를 사용하였다. 홈팀 순위 또한 상대 팀 순위와 동일하게 해당일 이전까지 홈팀의 순위를 사용하였다. 전처리 과정을 거친 데이터는 총 287개의 행으로 구성되었다.

3. 모델링

수집과 전처리 과정을 거친 데이터를 광주 - 기아 챔피언스 필드 일별 관중 수 예측을 위한 딥러닝 기반 선형 회귀 모델에 적용하였다. 모델링을 진행하기 위해 전처리 한 총 287개의 데이터 셋을 훈련 데이터 셋과 테스트 데이터 셋으로 나누었고, 각각 80%와 20%의 비율로 나누어 구성하였다.

본 연구는입력층과 6개의 은닉층, 출력층으로 구성된 딥러닝 모델을 설계하였다. 입력층을 통해 입력된 데이터는 256개, 128개, 64개, 32개, 16개, 8개의 노드를 가진 은닉층을 거친다. 출력층에서는 1개의 노드로 일별 관중 수 예측값을 출력한다. 딥러닝 기반 모델 구성 시, 과적합(Over-fitting)을 방지하기 위해 Dropout 레이어를 추가하였다. 첫 번째, 두 번째, 네 번째 은닉층은 Dropout의 비율을 20%로 설정하고, 세 번째 은닉층은 Dropout의 비율을 40%으로 설정하였다. 이를 통해 은닉층은 204개, 102개, 38개, 25개, 16개, 8개의 노드 정보가 전달된다. 각 은닉층의 활성화 함수(Activation Function)는 ReLU(Rectified Linear Unit)를 사용하였다. 출력층의 활성화 함수는 연속적인 값인 관중 수를 예측하기 위해 선형함수(Linear function)를 사용하였다. 손실함수(Loss Function)로는 MSE(Mean Squared Error)를 사용하였다. 최적화 함수(Activation Funchion)는 Adam Optimizer를 사용하였고, Epoch을 100으로 설정하여 총 100번의 반복 학습을 수행하도록 하였다. 제안하는 모델의 학습률(Learning Rate)은 0.01이며, 배치 크기(Batch Size)는 5를 사용하였다. 딥러닝 모델을 안정적인 성능을 검증하기 위해 10-Fold의 교차 검증(Cross Validation)을 진행하였다. 설계한 모델의 구성 요소와 하이퍼파라미터(Hyper-parameter)는 표 1과 같다.

표 1. 딥러닝 모델의 하이퍼파라미터
Table 1. Hyper-parameter of Deep Learning Model

Hyper-parameter	Value
은닉층 개수	6
은닉층 노드 수	204, 102, 38, 25, 16, 8
각 은닉층 별 Dropout 비율	20%, 20%, 40%, 20%, 0%, 0%
손실 함수	Mean Squared Error
최적화 함수	Adam Optimizer
Epoch	100
학습률	0.01
배치 크기	5

IV. 결과 및 검증

본 연구에서 설계한 딥러닝 모델의 우수성을 증명하기 위해 머신러닝 모델 중 연속형 변수를 예측하는 모델인 다중 선형 회귀 모델[12]을 사용하였다. 다중 선형 회귀 모델은 한 개의 종속 변수와 둘 이상의 여러 독립 변수 간의 관계를 선형 관계식을 통해 모델링하고, 종속 변수를 예측하는 모델이다. 모델링을 통해 각 독립변수가 종속 변수에 얼마나 영향을 미치는가를 수치로 표현할 수 있다는 장점이 있다. 하지만 노이즈(Noise) 데이터에 민감하다는 단점이 존재한다. 연구에서 연속형 변수를 예측하는 딥러닝 모델을 설계하였기 때문에 연속형 변수를 예측할 수 있는 다중 선형 회귀 모델을 비교 모델로 선정하고, 성능을 비교하였다.표 2는 제안한 딥러닝 모델과 머신러닝 모델의 성능 지표이다.

표 2. 일별 관중 수 예측 딥러닝, 머신러닝 모델의 정확도(Accuracy)
Table 2. Prediction Accuracy of Daily Attendance Using Deep Learning and Machine Learning Model

Model	Datasets		
	Training	10-Fold Cross Validation	Test
Deep Learning	**90.47%**	**90.46%**	**88.14%**
Machine Learning	88.1%	88.07%	86.49%

본 연구에서 제안한 딥러닝 모델이 훈련 정확도 90.47%, 검증 정확도 90.46%, 테스트 정확도 88.14%로 모든 데이터 셋에서 안정적으로 높은 성능을 보이는 것을 알 수 있다. 다중 선형 회귀 모델은 훈련 정확도 88.1%, 테스트 정확도 86.49%로 본 연구에서 설계한 딥러닝 기반 선형 회귀 모델보다 낮은 성능을 보였다. 이는 일별 관중 수를 예측하기 위해 사용한 각 독립 변수의 복잡한 특징을 딥러닝의 많은 파라미터를 활용하여 학습할 필요가 있음을 나타낸다.

V. 결 론

본 연구는 딥러닝(Deep Learning)을 활용하여 광주 - 기아 챔피언스 필드의 일별 관중 수를 예측하는 모델을 설계하였다. 코로나로 인해 무관중으로 경기를 진행한 2020년과 2021년을 제외하고, 2018년부터 2023년까지의 일별 기아 챔피언스 필드 관중 수를 종속변수로 사용하였다. 독립변수로는 시간, 날씨, 팀 기반 요소를 채택하여 딥러닝 및 머신러닝(Machine Learning) 기반 회귀(Regression) 모델로 모델링하였다. 실험 결과, 광주 - 기아 챔피언스 필드의 일별 관중 수를 예측한 딥러닝 모델이 머신러닝 모델보다 더 높은 성능을 가짐을 확인하였다. 본 모델을 통해 일별 관중 수를 예측함으로써 기아 타이거즈 구장의 마케팅 및 구장 내 입점 되어 있는 부대 시설의 재고 관리에 활용될 수 있을 것이라 기대한다. 그러나 본 연구에서 사용한 4개년도의 데이터가 딥러닝 학습을 진행하기에 충분히 크지 않다는 한계가 있다. 더 많은 데이터를 수집해 모델에 활용하는 경우 더욱 개선된 성능을 보일 것으로 판단한다.향후 연구로는 광주 - 기아 챔피언스 필드 뿐만 아니라 다른 팀 구장의 관중 수를 예측하고자 한다.

References

[1] J. Lee, "A Study on Determinants in Korean Pro-Baseball Spectators," *Journal of the Korean Data Analysis Society*, Vol.12 No.6, pp. 3507-3517, December 2010.

[2] J. Lee, "The Influence of Factors Affecting Decision to Spectate on Spectator Satisfaction and Revisiting Intention in Professional Baseball Games," *Korean Journal of Sport Management*, Vol.17 No.3, pp. 41-53, June 2012.

[3] Prosports Data Portal, http://data.prosports.or.kr

[4] Korean Baseball Organization, https://www.koreabaseball.com

[5] J. Park and S.H. Park, "A Study on Prediction of Attendance in Korean Baseball League Using Artificial Neural Network," *KIPS Tr. Software and Data Eng*, Vol. 6, No. 12 pp. 565~572 August 2017. DOI: 10.3745/KTSDE.2017.6.12.565

[6] J. Chea, "Prediction Model for Korean Professional Baseball Spectators," *Korean Journal of Sport Science,* Vol. 23, No. 4, pp. 892-905, December 2012.

[7] S. Lee, "Deep Structured Learning: Architectures and Applications," *The International Journal of Advanced Culture Technology(IJACT)*, Vol. 6, No. 4, pp. 262-265, 2018. DOI:10.17703//IJACT2018.6.4.262

[8] S. Oh, and M. Park, "Deep Learning-based Happiness Index Model Considering Social Variables and Individual Emotional Index," The Journal of the Convergence on Culture

Technology (JCCT), Vol. 10, No. 1, January 2024.
[9] H. Kim, "Study on the Prediction of the Number of Spectators and It's Factors in Pro Sports by Machine Learning Method," *Journal of the Korean Data Analysis Society*, Vol. 21, No. 4, pp. 1867-1880, August 2019.
[10] J. Cho and B. Seok, "The Development prediction model of Korea Professional Baseball League spectator using machine learning," *The Korea Journal of Sports Science*, Vol. 32, No. 5 pp. 547-558, October 2023. DOI: 10.35159/kjss.2023.10.32.5.547
[11] S. Nam and K. Jeon, "A Study on the Impact of Air Pollution on the Korean Baseball Attendance," *Korean Journal of Business Administration*, Vol. 32, No. 1, pp. 71-88, January 2019. DOI: 10.18032/kaaba.2019.32.1.71
[12] N. Ryu, H. Kim, and P. Kang, "Evaluating Variable Selection Techniques for Multivariate Linear Regression," Journal of the Korean Institute of Industrial Engineers, Vol. 42, No. 5, pp. 314-326, October 2016. DOI: 10.7232/JKIIE.2016.42.5.314

※ 이 논문은 서울여자대학교 학술연구비의 지원에 의한 것임 (2024-0026).