Korean Journal of Sport Science 2021, 32(3), 411-418
https://doi.org/10.24985/kjss.2021.32.3.411

Original Article

A Study on the Estimation Model for the Visitors to Let's Run Park Using Machine Learning

Jin Kook Kim*

Department of Sport and Healthcare, NamSeoul University, Chungnam, Korea

Article Info

Received 2021.07.12.
Revised 2021.09.22.
Accepted 2021.09.30.

Correspondence*

Jin Kook Kim
navyjk@daum.net

Key Words

Demand forecasting model,
Let's run park(Seoul),
Visitor,
Machine learning

PURPOSE The purpose of this study is to find the best model to predict the demand of visitors in Let's Run Park by using machine learning and to provide effective data for establishing future marketing strategies. **METHODS** For this purpose, three methods of machine learning were applied: random forest, adaboost, and gradient boosting. The variables for predicting the audience were weather data and the number of visitors per date for four years as training data, and the accuracy was predicted by comparing the actual data for one year. **RESULTS** First, the performance evaluation using random forest was conducted, RMSE = 1856.067, R^2 = .965, and error was 6.47%. Second, the performance evaluation using Adaboost was conducted, RMSE = 1836.227, R^2 = .965, and error was 5.25%, which was the lowest among the three machine learnings. Third, the performance evaluation using gradient boosting showed that RMSE = 1797.400 and R^2 = .967 were the most accurate among the three machine learnings and error was 6.99%. **CONCLUSIONS** As a result of this study, each of the three machine learning features existed, but the most efficient model was gradient boosting. In addition, the best way to utilize it in the field is to predict the number of visitors by comprehensively judging the results of the three machine learning, and it is judged that it will help efficient management decision making in the future.

서론

렛츠런 파크는 1988년 한국마사회가 주관하여 만든 경마장으로 수용인원 3만 5천명의 규모를 가지고 있고, 서울올림픽때 승마경기가 열린 국제규모의 경기장이다. 2014년 서울 경마공원을 '렛츠런 파크(서울)'로 브랜드 통합하여 개칭하였으며, 경마장뿐만 아니라 놀이공간, 공원, 마사 박물관, 체험 및 견학프로그램, 시크릿웨이 투어 등을 갖추고 있어 시민들의 가족 공원으로 건전한 레저 및 여가문화를 선도할 목적으로 운영되고 있다(Korea Racing Authority, 2021). 이러한 시민들의 레저공간으로 활용되고 있는 렛츠런 파크는 한국마사회가 운영하고 있는데, 무자본 특수법인의 형태로 수익금을 축산발전과 농어촌 복지기금으로 주로 사용하고 있다. 그렇기 때문에 렛츠런 파크는 본연의 목적을 달성하기 위해서 보다 효율적인 경영 전략을 가지고 운영할 필요가 있다.

이러한 측면에서 렛츠런 파크에 방문하는 입장객 수를 예측할 수 있다는 점은 경영학적인 관점에서 매우 중요하다. 우선 입장객 수를 통해 기본 입장료의 수준은 물론 여러 가지 프로그램에 대한 컨텐츠를 통해 다양한 수입원 창출이 가능해 지고, 또한 미래 지향적인 다양한 사업 발굴과 육성도 가능케 해 준다(Chiu, 2002). 즉, 입장객 수에 따른 재고관리는 물론 마케팅 및 예산 전략수립에 기본적으로 활용이 가능하다는 의미이다.

특히 거시경제 상황이 급변하거나 소비트렌드가 크게 변화하는 시점 그리고 대체기술의 부상 등 불확실성이 큰 경영환경에서는 더욱 더 수요예측에 대한 중요성이 강조된다(Park, 2012). 하지만 부적절한 수요 예측은 과잉투자나 기회손실 등의 경제적 문제뿐만 아니라 조직의 리더십 약화나 새로운 경쟁자의 출현 등의 훨씬 더 심각한 문제를 초래할 수도 있다. 때문에 이러한 수요 예측은 과거부터 경영진의 주요 관심사 중 하나였다. 과거에는 시계열 데이터를 주로 ARIMA 모델을 기반으로 수요를 예측했지만(Chae, 2012), 보다 정확한 예측을 위해 기계학습을 적용하거나 빅데이터를 활용하는 추세에 있다(Kim & Hong, 2020).

한편 수요 예측에 있어 중요한 단서는 수요에 영향을 주는 영향 변수를 적절하게 설정하고, 정확한 데이터를 수집하는 것이다(Hong, 2020). 렛츠런 파크 수요에 영향을 줄 수 있는 변수는 경마일정, 날씨, 이벤트 등이라고 볼 수 있는데, 본 연구에서는 경마 이용자가 대상이 아니기 때문에 경마일정에 대한 부분을 특별히 변수로 설정할 필요는 없고, 날씨와 요일 그리고 공휴일 유무를 통해서 몇 년간의 데이터를 축적하여 머신러닝을 활용하면 예측이 가능할 것으로 판단되었다.

유사한 연구로 프로야구 관람객 수요를 예측한 Yoo(2019)는 3년간 요일, 공휴일 유무, 상대 팀, 승패, 관중 수, 예매 현황 등을 기반으로 관중 패턴을 예측하였는데, 학습데이터의 양이 적어 오버 또는 언더피팅 현상이 발생한 부분을 지적하였다. 또한 최근 Park & Paik(2020)은 머신 러닝 기법을 활용해 상영관수, 주연배우, 상영시간, 평가자수, 구글관심도 등을 변수로 박스오피스 관람객을 예측했는데, 다양한 플랫폼의 데이터 크롤링을 통해 데이터를 추가해야 예측력을 높일 수 있다고 하였다. 이에 본 연구에서는 예측력을 높이기 위해서 충분한 훈련데이터 확보를 우선하였다.

따라서 본 연구는 날씨와 렛츠런 파크(서울)의 공공데이터를 활용하여 요일별 입장객 수를 예측하는 최적화 모델을 제시하는데 그 목적이 있다. 이를 통해 렛츠런 파크의 마케팅 전략 수립 시 미래 예측 가능한 날씨와 요일에 따른 입장객 수요를 예측하는데 도움이 될 것이다.

연구방법

연구대상 및 자료수집

본 연구의 대상은 한국마사회가 운영하고 있는 렛츠런파크(서울)이고, 날씨와 관련된 데이터는 공공데이터포털(https://www.data.go.kr)에서 제공하고 있는 날씨 정보를 2015년 1월 1일부터 2019년 12월 31일까지 과천지역의 일일 데이터를 제공받았다. 또한 렛츠런파크(서울)의 일일 입장객 데이터는 한국마사회가 운영하고 있는 공공데이터 제공 담당자를 통해 입수하여 분석에 활용하였다. 날씨 데이터와 동일하게 2015년부터 2019년까지 렛츠런파크(서울) 운영일인 금, 토, 일요일의 입장객 현황을 정제하여 이용하였다. 아울러 일별 요일과 공휴일에 대한 정보는 정부에서 제공하고 있는 표준 달력을 기준으로 지정하였다.

날씨 데이터의 경우 변수로 사용한 요인은 일별 최저 및 최고 기온, 일강수량, 최대풍속, 합계 일조시간, 일 최신적설, 평균 지면온도, 평균 상대습도, 최소 상대습도, 미세먼지 등이다. 렛츠런파크(서울)은 입장료가 2,000원으로 책정되어 있지만, 연간 몇 일 정도는 무료로 입장하는 경우가 있는데 이는 고려하지 않았다.

선정된 변수들의 5년 간 데이터를 정제한 후 2015년부터 2018년까지 4년간의 데이터는 머신러닝을 이용하여 학습을 수행하는 훈련용 데이터로 활용하였고, 2019년의 데이터는 테스트용 데이터로 사용하여 1년 동안 실제 입장객의 예측된 수와 실제 입장객의 수를 비교하였다. 코로나-19가 2020년 1월부터 시작되어 정상적인 영업활동이 어려운 부분을 감안하여 그 이전 5년 동안의 자료를 바탕으로 변수를 설정하였다.

Table 1. Input and output variables

Variables	Parameters	
Input variable	Date	the date, day, holiday
	Climatic	The minimum and maximum temperature (°C), daily precipitation (mm), maximum wind speed (ms), total sunshine time (hr), day deepest snow (cm), average ground temperature (°C), average relative humidity (%), minimum relative humidity (%), fine dust (μg/m^3)
Output variable	Let's Run Park(Seoul) number of daily visitors	

예측방법론

렛츠런파크(서울)의 입장객 수를 예측하기 위해 사용된 머신러닝의 종류로 본 연구에서는 랜덤포레스트, 에이다부스트 그리고 그래디언트 부스팅을 선정하여 진행하였다. 각 머신러닝의 종류는 장, 단점을 가지고 있어 다양한 방법의 기계학습을 통해 예측 결과의 정확도가 높은 방법을 적용하기 위해 3가지 방법을 이용하였다. 각 머신러닝의 특성과 분석 방법에 대한 원리는 다음과 같다.

1. 랜덤포레스트(random forest)

랜덤포레스트는 객관적으로 사실관계를 파악할 수 있는 결과변수에 대한 예측과 예측을 결정하는 요인을 탐색하기 위한 사회과학 연구에 활용되기 시작하였다(Choi & Min, 2018). 랜덤포레스트의 동작 원리는 주어진 훈련 데이터 집합으로부터 복원추출(random with replacement)을 하여 여러 개의 서로 다른 훈련 데이터 집한을 만들어 내는 기법을 붓스트랩(bootstrap)이라고 한다. 배깅(bagging)은 붓스트랩을 통해 여러 개의 훈련 데이터 집합을 만들고, 각 훈련 데이터 집합별로 분류기를 만들어 이들이 투표나 가중치 투표를 하여 최종 판정을 하는 방법이다. 훈련 데이터로부터 붓스트랩을 통해 3개의 훈련 데이터 집합을 만들고, 각 데이터 집합에 대하여 학습 알고리즘을 적용해 선형 분류기 3개를 만들어 이를 결합하여 분류하는 방식이다. 분류기로 결정트리를 사용하는 배깅 기법으로 대표적인 방법인 랜덤포레스트(random forest) 알고리즘이 있다. 랜덤포레스트는 모형별로 표본 선정 및 변수 선택에 있어 무작위성을 최대한으로 부여함으로써 독립적인 의사결정나무를 반복적으로 만들기 때문에 의사결정나무의 낮은 편향을 유지하면서도 분산을 낮추어 예측 오차를 줄일 수 있다(Géron, 2017). 또한 다수의 설명변수를 포함하는 고차원의 자료에서도 설명변수 간의 상호작용과 비선형성을 고려하므로 오류를 야기하지 않고 안정적이다. 따라서 예측의 정확도를 개선하여 일반화 가능성을 향상시켰다(Dangeti, 2017).

즉, 이 알고리즘은 붓스트랩을 통해 여러 개의 훈련 데이터 집합을 만들어 결정트리를 만든다. 결정트리를 만들어 가는 과정에서 분할 속성을 결정할 때, 모든 가능한 분할 속성을 고려하지는 않는다. 대신 분할 속성 후보들을 무작위로 선택한 다음, 이들에 대해서만 정보 이득 등을 계산하여 분할 속성과 분할 기준을 결정한다. 따라서 생성되는 각 결정트리는 서로 다른 형태가 될 수 있고, 앙상블 분류기를 만들 때 각 분류기가 특징 공간의 모든 속성을 고려하는 것이 아니라 서론 다른 소속성인 부분공간만을 사용하도록 하여 서로 다른 분류기가 만들어지도록 하는 부분 공간(subspace method) 방법을 적용한다(Lee, 2019). 랜덤포레스트의 경우 전처리가 거의 필요 없고, 매개변수가 없어도 기본적으로 좋은 성능을 만들기 때문에(Jung & Min, 2013) 본 연구에서 채택하였다.

2. 에이다부스트(Adaboost)

부스팅(boosting) 알고리즘은 여러 개의 분류기를 순차적으로 만들어 가는 앙상블 분류기 생성 방법이다. 이 알고리즘에서는 붓스트랩 기법으로 여러 훈련 데이터 집합을 만드는 것이 아니라, 각 학습 데이터의 가중치를 변경해 가면서 분류기를 만든다. 이때 가중치는 훈련 데이터 각각에 대한 오류를 계산하여 결정된다. 부스팅은 이전 단계의 분류기에서 잘못 분류된 데이터의 가중치는 높이고, 제대로 분류된 데이터의 가중치를 낮춘다. 부스팅 알고리즘의 동작과정을 보면 처음에는 모든 데이터가 같은 크기의 가중치를 갖고, 어떤 훈련 알고리즘을 적용하여 만들어진 분류기1이 몇 개의 데이터를 잘못 분류하면 두 번째 훈련 데이터 집합에 대해 만들어진 분류기2에서 가중치가 증가한다. 마찬가지로 분류기2에

서 잘못 분류된 데이터는 다음 분류기에서 가중치를 증가하는 방식이다. 이렇게 일련의 분류기를 훈련시키고, 각 분류기의 정확도를 측정한다. 새로운 질의 데이터가 주어지면 각 분류기가 분류 정보를 제공하고, 최종 분류 정보는 각 분류기의 정확도를 가중치로 하여 분류 결과를 결합하여 결정한다.

대표적인 부스팅 방법으로 에이다부스트(Adaboost) 알고리즘이 있는데, 이는 N개의 훈련 데이터가 있을 때 각 훈련 데이터 di 의 초기 가중치 ωi 로써 1/N을 부여한다. 따라서 전체 가중치의 합은 1이고, 훈련 오류값은 잘못 분류한 훈련데이터의 가중치의 분류기가 학습되면, 해당 분류기가 잘못 판정한 훈련 데이터의 가중치는 올리고, 제대로 판정한 훈련 데이터의 가중치는 내린다. 훈련 오류값이 ε라고 가정할 때, 분류기의 신뢰도 $\alpha = 0.5 \ln((1-\varepsilon)/\varepsilon)$로 정의한다. 그런 후 잘못 판정한 데이터 di 의 가중치 $\omega i = \omega i e^{\alpha}$ 로 변경하고, 제대로 판정한 데이터 dj 의 가중치 $\omega j = \omega j e^{-\alpha}$ 로 변경한다. 가중치를 변경한 후 가중치의 합이 1이 되도록 가중치 전체 합으로 각 가중치의 값을 나눈다. 이렇게 수정된 가중치를 갖는 훈련 데이터에 대하여 다시 분류기를 생성하는 과정을 반복한다. 학습 오류값이 0.5미만인 분류기들로 앙상블 분류기를 구성하면 높은 성능을 얻을 수 있다. 학습된 분류기들로 데이터의 부류를 판정할 때는 각 분류기의 신뢰도 α를 가중치로 하여 각 분류기의 판정 결과를 가중 다수결 투표(weighted majority voting) 방식으로 결합하여 최종 부류를 결정한다(Lee, 2019).

Adaboost 모형은 일부 변수만을 사용한 모델을 구현하므로, 타 머신러닝 알고리즘에 비해 데이터의 특성을 100% 반영하여 예측 성능이 저하되는 과적합(overfitting) 현상이 적게 발생하고, 신규 데이터에 대한 일반화(generalization)가 잘 이루어지는 장점이 있다. 본 연구에서 입장객 수를 예측하는 모형 수립이 필요하고, 분석 설계 특성상 표본 수가 많지 않은 점을 감안하여, Adaboost 알고리즘을 채택하였다(Raul, 2009).

3. 그래디언트 부스팅(gradient boosting)

부스팅 학습방법은 여러 개의 간단한 모델을 사용하거나 기본 학습기 혹은 성능이 약한 학습기를 여러 개 연결하여 성능이 높은 학습기를 만드는 앙상블 학습 방법으로 더 나은 학습 모델로 진행시켜 학습하는 방법이다. 대표적인 방법으로 앞서 설명한 에이다부스트와 그래디언트 부스팅이 있는데, 에이다부스트는 전체 학습 데이터셋을 이용하여 첫 모델을 만든 후 상대적으로 갖중치를 높인 후 두 번째 모델에 대하여 이 가중치로 업데이트하여 모델을 학습해 가면서 학습이 어려운 데이터에 대하여 적합(fit)하게 되도록 반복처리하는 방식이다. 반면에 그래디언트 부스팅은 데이터셋에 대하여 모델을 만든 후 모델의 오차(residual error)를 수정한 손실 함수(cost function)를 정의하고, 경사 하강법을 사용하여 다음에 추가될 트리가 예측해야 할 값을 보완하는 방법으로 모델을 추가해 주는 방식이다(Schapire, 2002).

이 방법에는 특성 중요도가 존재하며, 트리의 개수가 많아지는 경우 과적합이 될 수 있기 때문에 변수 튜닝이 필요하다. 고려할 매개 변수로서 학습율(learning rate)이 있다. 이것은 트리 오차를 얼마나 보완할 것인지를 나타내는 변수로써 값이 크면 복잡한 모델을 만든다. 그래디언트 부스팅은 매개 변수 튜닝이 필요하며, 트레이닝 시간이 길고, 트리 기반 모델의 특성으로 고차원 데이터셋에는 잘 작동하지 않는 단점이 존재한다(Woo et al., 2019). 그러나 주요 장점인 메모리를 적게 사용하면서도 빠른 예측이 가능한 이점이 있고, 다양한 분야에서 보편적으로 사용되는 모델이기 때문에(Kar et al., 2019) 본 연구에서도 채택하였다.

모델 설계

본 연구에서 적용한 머신러닝은 모두 Python 프로그램 환경에서 구축하였다. 알고리즘의 구축을 위해 설정되는 모델의 하이퍼 파라미터(hyper parameter)는 일반적으로 사용되는 수치를 준용하였고, 변수 간 상호작용의 깊이를 의미하는 depth는 2회로 설정하여 비교적 많지 않은 데이터의 크기로 인하여 발생할 수 있는 과적합의 문제를 저감하였다. 모형 추정에 앞서 훈련용 데이터(train data set)와 테스트용 데이터(test data set)을 분리하여 학습용 데이터를 추정하였으며, 두 가지 데이터 유형을 대상으로 한 모형의 정확도 차이가 크지 않아 과적합 문제는 발생하지 않은 것(Schapire & Singer, 1999)으로 판단하였다.

이를 기반으로 3가지 형태의 머신러닝에 적용된 데이터는 2015년부터 2018년까지 4년 간 데이터를 훈련용 데이터로 모형을 추정하였는데, 샘플링은 계층화된 폴드(stratified fold)의 수를 5로 지정하고, 10회 반복 및 전체 데이터의 70%의 수준으로 학습을 시켰다. 이후 2019년의 요일 및 날씨 정보(입력변수)만 제공하는 테스트용 데이터를 입력하여 입장객 수를 예측하여 나온 결과값과 실제 2019년 날짜별 입장객 수를 비교하여 머신러닝 종류별 예측 가능성 지표인 RMSE와 R^2의 값을 도출하였다.

연구결과

랜덤포레스트 모델 평가 및 예측 결과

랜덤포레스트를 이용하여 훈련용 데이터의 성능 평가를 한 결과, RMSE = 1856.067, R^2 = .965(96.5%)로 나와 매우 높은 정확도를 보였다. 랜덤포레스트 기법을 이용한 2019년 유료 입장객 수 예측 결과 렛츠런 파크(서울)는 금, 토, 일에만 영업을 하기 때문에 총 예측된 날은 150일(19년 1월 4일부터 12월 29일)이다. 예측된 입장객 수와 실제 입장객 수를 비교한 결과 150일 중 11일의 데이터 오차 범위가 커서 비정상 사례(abnormal case)가 발생하였고, 주요 일자별 입장객 차이는 아래의 〈Table 2〉와 같다. 전체적으로 2019년 실제 렛츠런 파크(서울) 유료 입장객은 총 3,098,340명이었고, 랜덤포레스트가 예측한 예상 입장객은 총 2,898,183명으로 오차는 6.47% 수준이었다.

아래의 〈Table 2〉에서 보이는 실제 입장객과 예측 입장객의 차이가

Table 2. Abnormal case of predictive results by random forest

Date	Day	Actual visitor	Predictive visitor	Deviation	Note
1.26	Sat	20,754	17,693	▲3,061	
1.27	Sun	31,316	27,341	▲3,975	
4.7	Sun	40,686	28,933	▲11,753	for free
4.13	Sat	26,377	19,620	▲6,757	festival
4.21	Sun	31,604	28,541	▲3,063	festival
8.3	Sat	17,759	21,577	▼3,818	36.0°C
8.4	Sun	23,665	28,342	▼4,677	34.4°C
9.8	Sun	40,103	28,647	▲11,456	for free
11.3	Sun	36,714	28,784	▲7,930	for free
12.8	Sun	37,050	27,126	▲9,924	for free
12.29	Sun	24,642	27,475	▼2,833	1.4mm

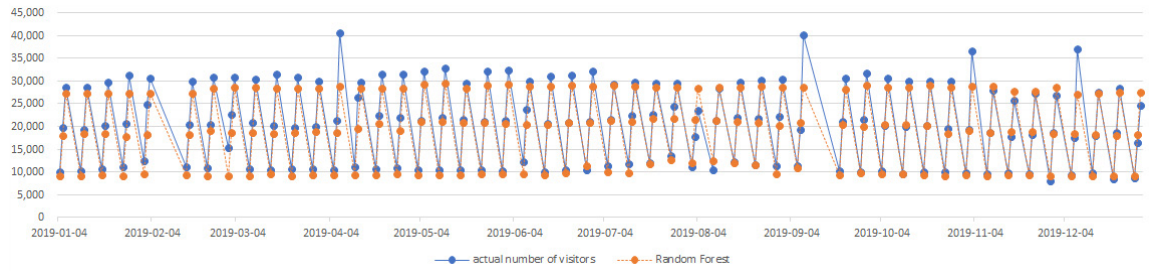

Fig. 1. Prediction of visitor patterns and actual number of visitors using Random forest

많이 나는 날은 4월 7일, 9월 8일, 11월 3일, 그리고 12월 8일이었고, 나머지 요일은 편차가 크지는 않았다. 전체적으로 편차가 크지 않은 날이 대부분이었으나, 편차가 큰 날은 '무료 입장' 이벤트를 하는 날이 3일이었고, 4월 13일에는 '서울 야간벚꽃 축제'가 있었고, 21일의 경우는 '플리마켓 축제'가 시행되어 오차가 발생한 것으로 판단된다. 8월에는 무더운 날씨로 인해 예측한 인원보다 적게 입장한 사례도 발견되었다.

한편 2019년 연간 실제 입장객 수와 예측 입장객 수의 분포는 위의 〈Fig. 1〉과 같다. 전체적인 분포는 매우 유사한 형태를 보이고 있으나, 전반적으로 예측 입장객 수가 실제 입장객 수 보다 조금 더 낮은 분포를 보이고 있어 언더피팅(underfitting)에 대한 정교한 정제 작업이 필요할 것으로 판단된다(Rathord et al., 2019).

에이다부스트 모델 평가 및 예측 결과

에이다부스트를 이용하여 훈련용 데이터의 성능 평가를 한 결과, RMSE = 1836.227, R^2 = .965(96.5%)로 나와 매우 높은 정확도를 보였다. 에이다부스트 기법을 이용한 2019년 유료 입장객 수 예측 결과 렛츠런 파크(서울)는 동일하게 총 예측된 날은 150일이다. 예측된 입장객 수와 실제 입장객 수를 비교한 결과 150일 중 16일의 데이터 오차 범위가 커서 비정상 사례(abnormal case)가 발생하였고, 주요 일자별 입장객 차이는 〈Table 3〉과 같다. 전체적으로 2019년 실제 렛츠런 파크(서울)의 유료 입장객은 총 3,098,340명이었고, 에이다부스트가 예측한 예상 입장객은 총 2,935,736명으로 오차는 5.25%로 3개의 머신러닝 중 가장 낮은 수준이었다.

Adaboost의 경우 랜덤포레스트와 달리 비정상 사례(abnormal case)가 더 많이 발생하였고, 정상 사례의 편차가 더 적은 것이 특징이었다. 즉, 예측이 비슷한 경우에는 편차가 매우 적었고, 예측의 범위를 벗어난 경우에는 편차가 랜덤포레스트에 비해 더 크게 나타난 것이다.

〈Table 3〉에서와 같이 2019년 초 겨울 시즌에 입장객에 대한 예측의 범위가 상대적으로 많이 벗어났다. 즉, 랜덤포레스트에 비하여 언더피팅(underfitting)에 대한 성향이 더욱 크게 나타나 실제 입장객이 더 많은 결과를 보이고 있다. 나머지 비정상 사례(abnormal case)는 랜덤포레스트와 유사한 분포를 보이고 있다.

Table 3. Abnormal case of predictive results by adaboost

Date	Day	Actual visitor	Predictive visitor	Deviation	Note
1.6	Sun	28,620	24,701	▲3,919	
1.20	Sun	29,788	24,701	▲5,087	
1.26	Sat	20,754	17,658	▲3,096	
1.27	Sun	31,316	27,330	▲3,986	
2.2	Sat	24,806	18,200	▲6,606	
2.17	Sun	29,972	24,701	▲5,271	
3.2	Sat	22,614	18,995	▲3,619	
4.7	Sun	40,686	30,138	▲10,548	for free
4.13	Sat	26,377	19,512	▲6,865	festival
4.21	Sun	31,604	28,672	▲1,932	festival
8.3	Sat	17,759	20,316	▼2,557	36.0°C
8.4	Sun	23,665	28,672	▼5,007	34.4°C
9.8	Sun	40,103	30,075	▲10,028	for free
11.3	Sun	36,714	30,075	▲6,639	for free
12.8	Sun	37,050	24,701	▲12,349	for free
12.29	Sun	24,642	27,330	▼2,688	1.4mm

그래디언트부스팅 모델 평가 및 예측 결과

그래디언트 부스팅을 이용하여 훈련용 데이터의 성능 평가를 한 결과, RMSE = 1797.400, R^2 = .967(96.7%)로 나와 3개의 머신러닝 중 가장 높은 정확도를 보였다. 그래디언트 부스팅 기법을 이용한 2019년 유료 입장객 수 예측 결과 렛츠런 파크(서울) 총 예측된 날은 마찬가지로 150일이다. 예측된 입장객 수와 실제 입장객 수를 비교한 결과 150일 중 11일

Table 4. Abnormal case of predictive results by gradient boosting

Date	Day	Actual visitor	Predictive visitor	Deviation	Note
1.26	Sat	20,754	16,973	▲3,772	
1.27	Sun	31,316	27,053	▲4,263	
2.2	Sat	24,806	17,069	▲7,737	
4.7	Sun	40,686	30,007	▲10,679	for free
4.13	Sat	26,377	18,955	▲7,422	festival
8.3	Sat	17,759	21,001	▼3,242	36.0°C
8.4	Sun	23,665	30,852	▼7,187	34.4°C
9.8	Sun	40,103	32,041	▲8,062	for free
11.3	Sun	36,714	29,780	▲6,934	for free
12.8	Sun	37,050	26,806	▲10,244	for free
12.29	Sun	24,642	27,266	▼2,624	1.4mm

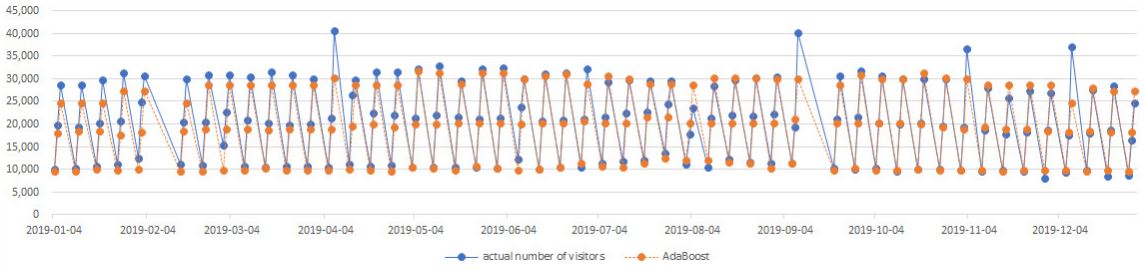

Fig. 2. Prediction of visitor patterns and actual number of visitors using Adaboost

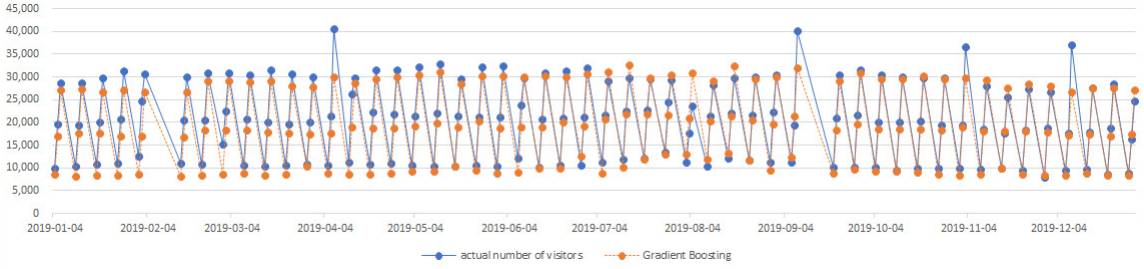

Fig. 3. Prediction of visitor patterns and actual number of visitors using Gradient boosting

의 데이터 오차 범위가 커서 비정상 사례(abnormal case)가 발생하였고, 주요 일자별 입장객 차이는 〈Table 4〉와 같다. 전체적으로 2019년 실제 렛츠런 파크(서울) 유료 입장객은 총 3,098,340명이었고, 그래디언트 부스팅이 예측한 예상 입장객은 총 2,882,068명으로 오차는 6.99% 수준이었다.

Gradient boosting의 경우 랜덤포레스트와 매우 유사한 패턴의 비정상 사례(abnormal case)가 발생하였고, 정상 사례의 편차는 더 적은 것이 특징이었기 때문에 3개의 머신러닝 중 가장 높은 정확도를 보이는 것으로 나타났다. 즉, 예측이 비슷한 경우와 예측의 범위를 벗어난 경우에도 편차가 랜덤포레스트에 비해 더 적게 나타난 것이다. 〈Table 4〉에서와 같이 랜덤포레스트와 동일한 날에 편차가 크게 나타났고, 전체적으로 편차의 범위가 랜덤포레스트보다 낮은 경향을 보였다. 그러나 전체 패턴에 대한 부분은 다른 머신러닝 처럼 언더피팅(underfitting)에 대한 성향이 더욱 크게 나타나 실제 입장객이 더 많은 결과를 보이고 있다.

그러나 일부 날짜에 대한 예측의 차이가 존재하였는데, 2월 2일의 경우 랜덤포레스트는 정상 범위에서 예측을 하였지만, 그래디언트 부스팅은 특별한 이유없이 편차가 크게 예측되었다. 하지만 4월 21일 '플리 마켓' 이벤트가 있어 랜덤포레스트의 경우 편차가 크게 나타났지만, 그래디언트 부스팅의 결과는 정상 범위 안에서 예측이 되었다. 이렇듯 머신러닝별 데이터 훈련의 방식의 차이가 존재하기 때문에 동일한 데이터를 가지고 훈련을 하더라도 결과 예측에 대한 값은 다르게 나타나고 있다. 전체적으로 일요일 입장객 수에 대한 편차가 공통적으로 많이 나타난 점도 고려해야 할 부분이다.

논의

지금까지의 결과를 바탕으로 한 논의는 다음과 같다. 렛츠런 파크(서울)의 입장객 예측을 위한 머신러닝 별 특징을 보면 3개의 정확도는 96.5~96.7%로 매우 높은 정확도를 보여 날씨와 일자에 따른 입장객 예측이 가능한 것으로 보였다. 그러나 무료입장 이벤트나 특정 시즌에 펼쳐지는 특별 이벤트가 있는 날에는 머신러닝의 예측율이 떨어지는 것으로 보아 훈련용 데이터의 변수 설정 시 이벤트 유무에 대한 변수를 새롭게 설정하면 성능은 더욱 좋아질 것으로 판단된다. 또한 여름철과 겨울철에는 최고 및 최저 온도, 강수량 등의 변수에 따라 입장객 수의 차이가 나타나는 현상을 확인할 수 있었기 때문에 이 변수에 대한 전처리도 필요하다고 판단된다(Kim, 2019). 아울러 요일별 특성은 보면 금요일의 경우는 비정상 사례(abnormal case)가 나타나지 않았고, 일요일에 주로 많이 나타나는 현상이 발생하였기 때문에 변수 설정 시 별도의 전처리가 필요해 보인다.

한편 랜덤포레스트, 에이다부스트, 및 그래디언트 부스팅의 3가지 머신러닝별 특징은 전체적인 관점에서 유사한 성능을 보였지만, 세부적인 특징은 조금씩 다르게 나타났다. 우선 공통적으로 실제 입장객 보다 낮은 수치의 예측을 하고 있는 언더피팅이 많이 발생하였는데, 언더피팅(underfitting)의 경우 모델이 충분히 복잡하지 않아 훈련 데이터에 있는 패턴을 모두 잡아내지 못할 때 발생할 가능성이 크기 때문에 층의 개수나 유닛의 개수를 추가하여 복잡도가 더 높은 모델을 사용하거나 가중치의 규제를 완화하여 해결할 수 있다(Lee, 2019). 아울러 학습 데이터의 양이 적은 경우에도 이런 현상이 나타날 수 있기 때문에 데이터를 추가적으로 확보하여 진행하는 것도 해결 방법이라 하겠다(Yoo, 2019).

또한 세부적인 관점에서 랜덤 포레스트를 기준으로 다른 머신러닝과 비교하여 설명하면, 특히 에이다부스트의 경우 예측의 정확도가 높을 때는 거의 유사한 입장객을 예측하고 있지만, 정확도가 떨어질 경우 다른 머신러닝보다 편차가 크게 나타나고 있다. 그렇기 때문에 연간 실제 입장객과 예측 입장객의 편차가 가장 낮은 것으로 나타났고, 비정상 사례(abnormal case)는 가장 많이 발생하였다. 그래디언트 부스팅의 경우는 가장 높은 정확도를 보이고 있었지만, 특정 이벤트가 있거나 날씨 변수에 의한 비정상 사례(abnormal case)의 일자에는 랜덤포레스트보다

편차가 약간씩 더 크게 나타났다. 그러나 그래디언트 부스팅은 다양한 분야의 예측 연구에 활용되는 머신러닝으로 비교적 정확도 높은 것으로 평가받고 있기 때문에(Heo et al., 2018). 훈련용 데이터 전처리를 강화한다면 본 연구 대상의 예측에 있어서도 가장 적합할 것으로 판단된다.

그러므로 향후 렛츠런 파크 운영에 있어 예측할 수 있는 미래의 날씨와 요일 정보를 가지고 본 연구에서 적용한 머신러닝의 훈련 데이터 결과를 활용하면 적절한 시점에서 이벤트 개최 시 많은 입장객을 수용할 수 있고, 또한 예측된 입장객의 수요에 따라 마케팅 기획도 가능해질 수 있기 때문에 효율적이고, 효과적인 마케팅 의사결정이 가능할 수 있을 것으로 판단된다. 이렇듯 머신러닝은 종류에 따라 다양한 특징들을 가지고 있기 때문에(Yoo, 2019) 특정 머신러닝을 이용하여 입장객을 예측하기 보다는 여러 가지 머신러닝을 지속적으로 훈련시켜 각각의 장점을 발견하여 보다 정교한 모델을 찾아가는 노력이 필요해 보인다. 또한 본 연구에서 예측하지 못한 요인들을 추가로 고려하여 정제한다면 보다 높은 정확도를 가진 모형을 구축할 수 있기 때문에 이 부분에 대한 고려도 반드시 필요하다 하겠다.

결론 및 제언

본 연구는 랜덤포레스트, 에이다부스트 및 그래디언트 부스팅 머신러닝을 이용하여 날씨에 일자에 대한 변수를 바탕으로 렛츠런 파크(서울)의 입장객을 예측하는 최적의 모델을 구축하는데 목적이 있다. 이를 위해 2015년부터 2018년까지의 4년간 데이터를 훈련용 데이터로 사용하여 훈련시켰고, 2019년도 실제 입장객과 머신러닝별 예측 입장객 수를 비교하여 성능을 확인하였다. 연구목적에 따른 연구결과는 다음과 같다.

첫째, 랜덤포레스트를 이용하여 성능 평가를 실시한 결과 RMSE = 1856.067, R^2 = .965였고, 오차는 6.47% 이다. 둘째, 에이다부스트를 이용하여 성능 평가를 실시한 결과 RMSE = 1836.227, R^2 = .965였고, 오차는 5.25%로 3개의 머신러닝 중 가장 낮았다. 셋째, 그래디언트 부스팅을 이용하여 성능 평가를 실시한 결과 RMSE = 1797.400, R^2 = .967로 3개의 머신러닝 중 가장 정확도가 높았고, 오차는 6.99% 이다.

3개의 머신러닝은 각각의 특징이 존재하였으나, 가장 성능이 우수한 모델은 그래디언트 부스팅이었다. 또한 모든 머신러닝의 결과가 전반적으로 언더피트(underfitting)의 경향을 보여 보다 정교한 모델을 구출하기 위해서는 이벤트, 날씨 등의 변수에 대한 전처리가 더욱 요구된다고 하겠다. 이를 통해 보다 정확도가 높은 모델을 구축하면 실제 현장에서도 충분히 적용 가능할 것이다. 아울러 머신러닝을 활용하는 가장 좋은 방법은 3개의 머신러닝의 결과를 종합적으로 판단하여 입장객 수를 예측하는 것이고, 결과를 기반으로 렛츠런 파크(서울)의 효율적인 마케팅 의사결정에 도움을 줄 것으로 판단되고, 나아가 정부 유관기관의 정책수립에도 실효성 있는 기초자료로 활용될 수 있을 것으로 판단된다.

한편 본 연구를 진행하면서 향후 연구를 위한 제언을 보면, 보다 정교한 예측을 위해서는 경마 일정에 대한 변수 역시 예측변수라 판단이 되고, 요일별 또는 날씨별 머신러닝 종류에 따른 비교 연구를 진행하는 것도 후속연구로 진행될 필요가 있을 것으로 생각된다. 또한 시민들이 이용하는 공원이기 때문에 특별 이벤트에 대한 변수를 고려하여 머신러닝의 변수를 지정하면 수요예측에 대한 정교함은 더욱 높아질 수 있을 것이다.

참고문헌

Chea, J. S. (2012). Prediction model for Korean professional baseball spectators. *Korean Journal of Sport Science, 23*(4), 892-905

Chiu, C. (2002). A case-based customer classification approach for direct marketing. *Expert Systems with Applications, 22*, 163-168.

Choi, P. S., &, Min, I. S. (2018). A predictive model for the employment of college graduates using a machine learning approach. *Journal of Vocational Education & Training, 21*(1), 31-54.

Dangeti, P. (2017). *Statistics for machine learning*. Birmingham: Packt Publishing Ltd.

Géron, A. (2017). *Hands-on machine learning with scikit-learn and tensorflow, sebastopol*. CA: O'Reilly Media.

Heo, J. S., Kwon, D. H., Kim, J. B., Han, Y. H., & An, C. H. (2018). Prediction of cryptocurrency price trend using gradient boosting. *KIPS Transactions on Software and Data Engineering, 7*(10), 387-396.

Hong, K. H. (2020). A predictive model for suicidal ideation of adolescents using random forests machine learning algorithm. *Korean journal of social welfare, 72*(3), 157-180.

Jung, J. H., & Min, D. K. (2013). The study of foreign exchange trading revenue model using decision tree and gradient boosting. *Journal of the Korean Data And Information Science Society, 24*(1), 161-170.

Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2019). Predicting the direction of stock market price using tree based classifiers. *The North American Journal of Economics and Finance, 47*, 552-567.

Kim, H. (2019). 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, 21*(4), 1867-1880.

Kim, E. M., & Hong, T. H. (2020). A deep learning approach for the prediction of tourism demand using online review sentiment analysis. *Korea Society of IT Service 2020 spring conference*, 729-732.

Korea Racing Authority (2021). https://park.kra.co.kr/seoul_main.do

Lee, G. M. (2019). *Artificial intelligence*. Seoul: Saenung.

Park, D. K., & Paik, J. R. (2020). Prediction of movies box-office success using machine learning approaches. *Proceedings of the Korean Society of Computer Information Conference, 28*(1), 15-18.

Park, S. B. (2012). *Effective demand prediction methods and cases*. SERI Issue paper.

Raul, R. (2009). *AdaBoost and the super bowl of classifiers: A tutorial introduction to adaptive boosting*. Unpublished Master's Thesis, Freie University: Berlin.

Rathord, P., Jain, A., & Agrawal, C. (2019). A comprehensive review on online news popularity prediction using machine learning approach. *International Journal Online of Science, 5*(1), 1-29.

Schapire, R. E. (2002). *The boosting approach to machine learning, an overview*. Heidelberg: Springer.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine learning, 37*(3), 297–336.

Woo, Y. C., Lee, S. Y., Choi, W., Ahn, C. W., & Baek, O. K.(2019). Trend of utilization of machine learning technology for digital healthcare data analysis. *Electronics and Telecommunications Trends, 34*(1), 98-110.

Yoo, J. H. (2019). A Study on the prediction of the demand of the spectator using machine learning. *Journal of IEEE Korea Council, 23*(4), 128-134.

머신러닝을 활용한 렛츠런 파크 입장객 수요 예측 최적화 모델 연구

김진국

남서울대학교 겸임교수

[목적] 본 연구는 머신러닝을 활용하여 렛츠런 파크의 입장객 수요를 예측하는 최적의 모델을 발견하여 향후 마케팅 전략을 수립하는데 실효성 있는 자료를 제공하는데 그 목적이 있다.
[방법] 이를 위해 머신러닝 방법을 랜덤포레스트, 에이다부스트, 그래디언트부스틴의 3가지 방법을 적용하였고, 입장객 예측을 위한 변수는 날씨 데이터와 4년 간 날짜별 입장객 수를 훈련 데이터로 설정하고, 1년간 실제 데이터와 비교하여 정확도를 예측하였다.
[결과] 첫째, 랜덤포레스트를 이용하여 성능 평가를 실시한 결과 RMSE=1856.067, R^2=.965였고, 오차는 6.47% 이다. 둘째, 에이다부스트를 이용하여 성능 평가를 실시한 결과 RMSE=1836.227, R^2=.965였고, 오차는 5.25%로 3개의 머신러닝 중 가장 낮았다. 셋째, 그래디언트 부스팅을 이용하여 성능 평가를 실시한 결과 RMSE=1797.400, R^2=.967로 3개의 머신러닝 중 가장 정확도가 높았고, 오차는 6.99% 이다.
[결론] 본 연구의 결과 3개의 머신러닝은 각각의 특징이 존재하였으나, 가장 성능이 우수한 모델은 그래디언트 부스팅이었다. 또한 모든 머신러닝의 결과가 대부분 언더피트(underfitting)의 경향을 보여 보다 정교한 모델을 구출하기 위해서는 이벤트, 날씨 등의 변수에 대한 전처리가 더욱 요구된다고 하겠다. 아울러 현장에서 활용할 수 있는 가장 좋은 방법은 3개의 머신러닝의 결과를 종합적으로 판단하여 입장객 수를 예측하는 것이 가장 좋고, 이를 통해 효율적인 마케팅 의사결정에 도움을 줄 것으로 판단된다.

주요어
수요 예측 모델, 머신 러닝, 렛츠런 파크, 입장객 수요