어린이대공원 입장객 수 예측 모델 비교

2022104007 전지현

목차

1. 서론

서울어린이대공원은 도심 내에서 시민들이 자연을 즐기고 여가를 보낼 수 있는 주요 녹지 공간으로, 연중 다양한 계절적·기후적 요인에 따라 입장객 수가 뚜렷한 변동성을 보인다. 특히 방학과 주말에는 가족 단위 방문이 증가하며, 이로 인한 혼잡도는 방문객의 체류 만족도와 시설 운영 효율성 모두에 영향을 미칠 수 있다. 이에 따라 정확한 입장객 수 예측은 공원 운영 계획 수립, 인력 배치, 시설 관리 및 마케팅 전략 수립 등 여러 측면에서 실질적 효용을 제공할 수 있다.

본 연구는 이러한 문제의식 하에, 공공 데이터를 기반으로 서울어린이대공원의 일별 입장객 수를 예측할 수 있는 통합적 모델을 구축하는 것을 목적으로 한다. 예측에 활용된 데이터는 서울시 공공데이터, 기상청 종관기상관측 및 황사관측 자료, 교육부의 학사일정 정보 등 다원적 출처에서 수집되었으며, 이를 통합하여 다양한 외부 요인들이 입장객 수에 미치는 영향을 정량적으로 분석하였다. 2022년 1월부터 2024년 12월까지의 데이터를 기반으로, 시계열 기반 베이스라인 모델인 Naïve 및 Seasonal Naïve와 머신러닝 기반의 Random Forest 및 XGBoost 모델을 비교 분석하였다. 또한, 변수 중요도 분석 및 시계열적 파생 변수 생성을 통해 모델의 예측 성능과 해석력을 동시에 확보하고자 하였다.

2. 데이터 및 전처리

2.1 데이터 출처 및 변수 설명

본 연구에 사용된 데이터는 총 네 가지 범주의 데이터로 구성되며, 모든 데이터는 일 단위 관측값으로 정렬 및 병합되었다. 첫째, 입장객 수 데이터는 서울시 공공데이터포털에서 제공하는 "서울어린이대공원 일일입장객수" 자료로부터 수집되었으며, 변수는 날짜(date)와 해당 일자의 총 입장객 수(visitors_total)로 구성된다. 해당 변수는 유료 및 무료 입장객 수의 합산값으로 계산되었다. 둘째, 기상 데이터는 기상청 종관기상관측(ASOS) 자료를 기반으로 하며, 최고기온(tmax), 최저기온(tmin), 평균 상대습도(humidity), 일강수량(rain), 일 최심적설(snow), 최대 풍속(wind)으로 구성된 총 여섯 개의 연속형 변수로 구성되어 있다. 셋째, 대기질 데이터는 기상청 황사관측 자료에서 일별 미세먼지 농도(PM10, 변수명 air)를 추출하여 활용하였다. 넷째, 학사일정 및 요일 정보는 교육부 및 공공 달력 데이터를 기반으로 주말 여부(weekend)와 방학 여부(vacation)를 더미 변수 형태로 생성하였다. 최종적으로 모든 데이터를 날짜 기준으로 병합하여 총 1,096 일치의 관측값을 포함한 통합 데이터프레임을 구성하였다.

2.2 결측치 처리 및 이상치 처리

데이터 전처리 과정에서 발생한 결측치는 변수의 성격에 따라 적절히 처리하였다. 입장객 수 데이터의 결측치는 주로 공원 휴장일로 인한 것으로 판단되어 해당 일자의 값은 0으로 대체하였다. 기상 변수의 경우, rain 및 snow의 결측은 해당 일에 강수 및 적설이 없었던 것으로 간주하고 0으로 처리하였다. tmin의 결측은 측정 오류로 판단하였으며, 동일 일자의 최고기온 및 평균기온 차이 (당일 최고 기온 – (당일 최고 기온 - 당일 평균기온))를 바탕으로 추정하여 보간하였다. wind 변수의 결측은 인접한 전일의 측정값을 활용해 보완하였다. 모든 결측치 처리는 변수의 물리적 의미와 데이터 수집 방식의 특성을 고려하여 수행되었으며, 통계적 왜곡을 최소화하는 방향으로 처리하였다.

한편, 본 연구에서는 트리 기반 모델인 Random Forest와 XGBoost를 주된 예측 기법으로 사용하였기 때문에, 일반적인 회귀 모델에서 필요로 하는 정규화(스케일링) 및 다중공선성 제거 절차는 생략하였다. 트리 기반 모델은 변수 간 선형 관계보다는 분할 기준의 이질성에 기반하여 작동하기 때문에, 상관관계가 있는 변수가 공존하더라도 모델 성능에 악영향을 주지 않는 것으로 알려져 있다. 이에 따라 본 연구에서는 데이터의 원천적 분포를 유지한 상태에서 입력값을 그대로 모델에 적용하였다.

2.3 변수 생성

모델 성능 향상을 위하여 시계열 특성이 반영된 파생 변수들도 함께 생성하였다. 전일 방문객 수(lag_1), 7일 전 방문객 수(lag_7), 최근 7일간 이동 평균 방문객 수(rolling_7)는 각각 단기적 변동성과 주간 반복성을 반영하기 위한 변수들이다. 또한 날짜 변수로부터 요일(day_of_week), 월(month), 연도(year) 등의 시간 관련 파생 변수들을 도출하여 계절성 및 주기성을 반영하였다.

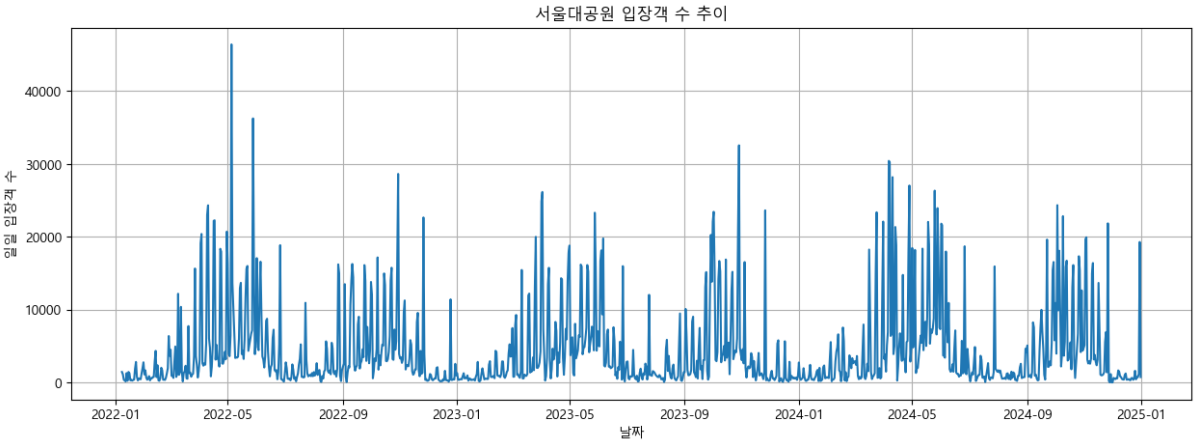
2.4 데이터 분할

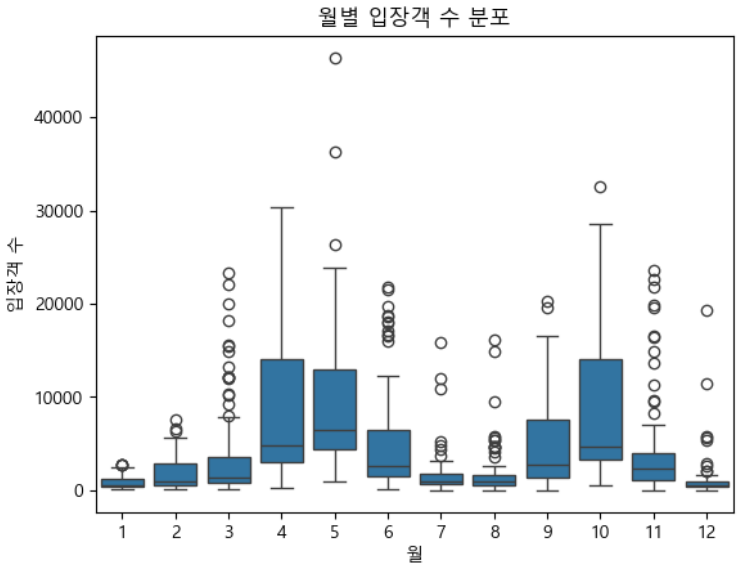
데이터는 시계열의 연속성을 고려하여 학습용, 검증용, 테스트용 세트로 분할되었다. 학습 데이터는 2022년 1월 1일부터 2023년 12월 31일까지, 검증 데이터는 2024년 1월 1일부터 6월 30일까지, 테스트 데이터는 2024년 7월 1일부터 12월 31일까지 구성되었으며, 하이퍼파라미터 튜닝 및 모델 선택은 검증 세트를 기준으로 수행되었다. 테스트 세트는 최종적인 모델 성능을 평가하는 데에만 활용되었으며, 데이터 누설을 방지하기 위한 시간 순서 기반의 분할 방식을 적용하였다.

2.5 EDA

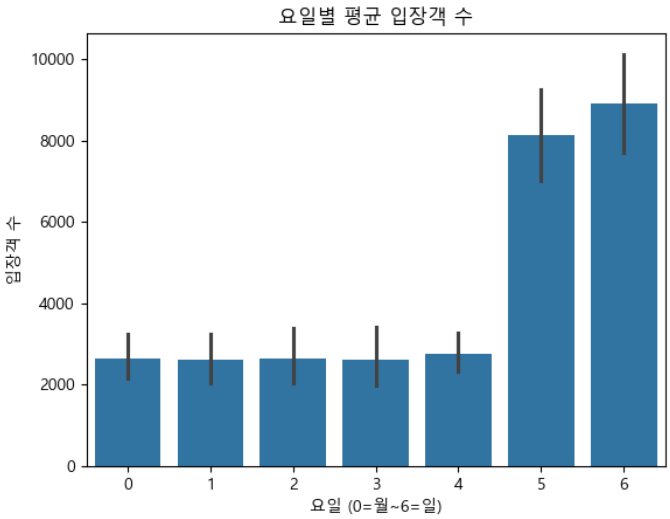
탐색적 데이터 분석(Exploratory Data Analysis, EDA)을 통해 서울대공원 입장객 수의 시계열적 특성과 외부 요인과의 관계를 다각도로 분석하였다. 먼저 전체 입장객 수의 시계열 추이를 살펴본 결과, 데이터는 뚜렷한 계절성과 주기성을 나타냈다.

월별 입장객 수의 분포를 boxplot으로 시각화한 결과, 4월과 5월, 그리고 10월과 11월에 방문객 수의 중앙값과 최대값이 모두 높게 나타났으며, 특히 5월과 10월에는 3만 명 이상에 달하는 이상값(outlier)들이 존재하였다. 이러한 결과는 봄과 가을의 기후 조건이 실외 활동에 유리하다는 점과 맞물리며, 계절성이 입장객 수 예측에 있어 중요한 변수로 작용할 수 있음을 뒷받침한다.

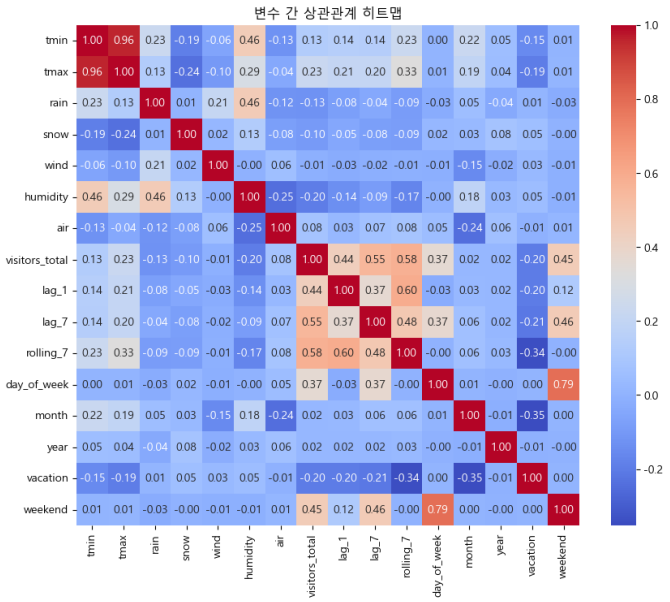




이와 함께, 요일별로도 뚜렷한 차이가 관측되었다. 요일을 기준으로 입장객 수의 평균을 비교한 결과, 주말인 토요일(5)과 일요일(6)에 약 8,000명 이상의 높은 수요가 나타나는 반면 평일은 평균 2,500~3,000명 내외의 비교적 낮은 방문객 수를 기록하였다. 이는 대공원이 전형적인 여가 공간으로 활용되고 있음을 시사하며, 주말과 평일 간 이용 행태에 명확한 차이가 존재함을 의미한다.



상관계수 행렬을 통해 변수 간 관계를 정량적으로 분석한 결과, 최저기온(tmin)과 최고기온(tmax) 간의 상관계수는 0.96으로 매우 강한 양의 상관성을 나타냈다. 이는 두 변수 간 중복성이 높음을 의미하며, 예측 모델 설계 시 해석력 중심의 모델을 구성할 경우 선택적으로 제외할 수 있는 후보가 된다. 또한 day_of_week와 weekend 변수 간에도 상관계수 0.79로 매우 높은 상관관계가 나타났는데, 이는 주말을 판단하는 이진 변수(weekend)가 요일을 정량화한 변수(day_of_week)로부터 도출되었기 때문으로 해석된다. 아울러 date와 year 변수 간에도 상관계수 0.94 이상으로 높은 시간 종속적 구조가 확인되었다. 분석 결과 일부 변수들 간에 높은 상관관계가 관찰되었으나, 본 연구에서는 트리 기반 모델이 변수 간 상관성에 민감하지 않다는 특성을 고려하여, 별도의 변수 제거 없이 전 변수 집합을 그대로 모델에 적용하였다.



3. 모델링

본 프로젝트에서는 시계열 특성을 가진 데이터를 활용하는 만큼, 베이스라인 모델과 함께 머신러닝 기반의 회귀 모델을 비교함으로써 예측 성능의 향상 가능성과 실무적 활용도를 함께 평가하고자 하였다.

3.1 베이스라인 모델

먼저, 비교 기준 역할을 수행할 베이스라인 모델로는 Naïve 모델과 Seasonal Naïve 모델을 사용하였다. Naïve 모델은 $\hat{y}_t = y_{t-1}$ 형태로, 직전 일자의 입장객 수를 그대로 다음 날의 예측값으로 사용하는 방식이다. 이는 어떠한 학습 과정 없이 이전 값에 전적으로 의존하는 매우 단순한 예측 방식이다. 반면 Seasonal Naïve 모델은 주기성을 반영하여 $\hat{y}_t = y_{t-7}$ 형태로 7일 전의 입장객 수를 그대로 예측값으로 활용한다. 이는 입장객 수가 주간 단위로 반복되는 패턴을 보일 가능성을 고려한 방식이다.

3.2 머신러닝 모델

본 연구의 주요 예측 기법으로는 머신러닝 기반의 Random Forest Regressor와 XGBoost Regressor를 채택하였다. Random Forest는 다수의 결정 트리를 생성하여 각기 다른 데이터 샘플과 피처 조합에 대해 독립적으로 학습한 후, 그 예측 결과를 평균 내는 방식으로 작동하는 앙상블 학습 기법이다. 트리 기반 모델은 변수 간의 선형 관계보다는 분할 기준의 비선형적 이질성을 중심으로 학습되기 때문에, 다양한 변수 간 상호작용을 효과적으로 반영할 수 있다는 장점을 가진다.

XGBoost는 Gradient Boosting 방식의 대표적인 구현체로, 각 반복 단계에서 이전 모델의 예측 오차를 보완해가며 모델을 점진적으로 개선하는 부스팅 기법이다. 학습 과정에서 손실 함수를 기반으로 모델을 최적화하며, 과적합을 방지하기 위한 정규화 항과 병렬 처리 구조를 통해 빠르고 안정적인 예측 성능을 확보할 수 있다.

두 모델의 하이퍼파라미터 최적화에는 Optuna 프레임워크가 사용되었다. Optuna는 베이지안 최적화 알고리즘에 기반하여 탐색 공간 내에서 효율적인 탐색을 수행하며, 기존의 Grid Search나 Random Search에 비해 적은 반복 횟수로도 더 우수한 성능을 얻을 수 있다. 본 연구에서는 Random Forest의 경우 n_estimators, max_depth 등의 파라미터를, XGBoost의 경우 learning_rate, n_estimators, max_depth 등의 파라미터를 대상으로 하여 검증 데이터셋의 RMSE를 최소화하는 조합을 탐색하였다.

3.3 성능 지표

모델의 성능 평가는 다음의 네 가지 지표를 기준으로 수행되었다. 첫째, RMSE(Root Mean Squared Error)는 예측값과 실제값의 차이를 제곱하여 평균한 후 제곱근을 취한 값으로, 큰 오차에 민감하게 반응하며 예측의 정밀도를 측정하는 기본적인 지표이다. 둘째,

MAE(Mean Absolute Error)는 절대 오차의 평균으로, 이상치에 덜 민감하며 평균적인 예측 오차 크기를 직관적으로 보여준다. 셋째, MAPE(Mean Absolute Percentage Error)는 예측 오차를 실제값 대비 백분율로 환산한 값으로, 상대적 예측 정확도를 평가할 수 있는 지표이나, 실제값이 0에 가까운 경우 값이 과도하게 커질 수 있다는 단점이 있다. 마지막으로, R^2(Coefficient of Determination)는 모델이 전체 데이터 분산을 얼마나 잘 설명하는지를 나타내며, 1에 가까울수록 높은 설명력을 가진다. R^2가 0보다 작을 경우, 해당 모델은 단순 평균보다도 성능이 떨어지는 것으로 간주된다.

4. 실험 및 결과

4.1 모델별 성능 비교

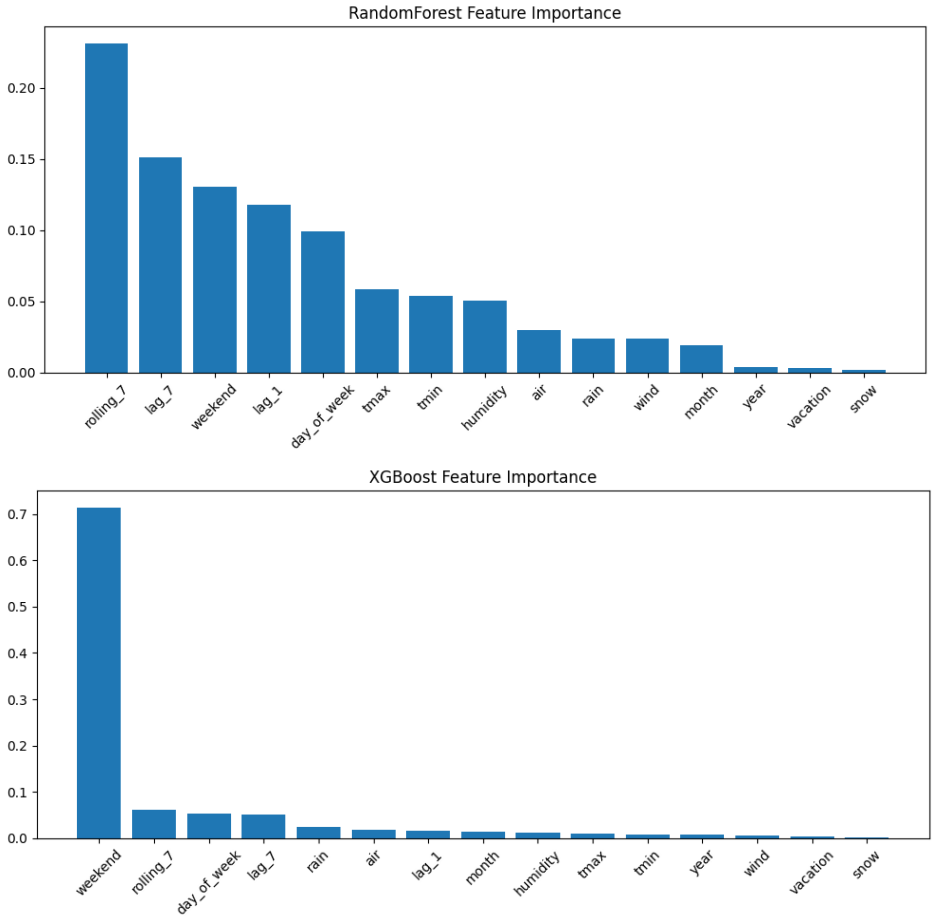
모델	RMSE	MAE	MAPE	R^2
Naïve	6046.60	3185.57	186.44%	-0.250
Seasonal Naïve	5228.03	2656.84	118.21%	0.089
Random Forest	3410.50	1672.95	88.81%	0.601
XGBoost	3387.67	1781.62	87.09%	0.606

가장 기본적인 Naïve 모델의 경우, RMSE 는 6,046.60, MAE 는 3,185.57, MAPE 는 186.44%, R^2는 -0.250 으로 나타났다. 이는 직전 일자의 수치만을 기반으로 예측하는 방식이 시계열의 계절성, 주말 효과, 외부 환경 변수 등을 고려하지 못하므로 실질적인 예측력은 거의 없음을 의미한다. Seasonal Naïve 모델은 주간 주기를 반영하여 약간의 성능 향상이 있었으며, RMSE 는 5,228.03, MAE 는 2,656.84, MAPE 는 118.21%, R^2는 0.089 로 측정되었다. 그러나 여전히 실무 적용에는 적절하지 않을 만큼 낮은 설명력을 보였다.

반면, 머신러닝 기반의 모델에서는 현저한 성능 향상이 관찰되었다. Random Forest 모델은 RMSE 3,410.50, MAE 1,672.95, MAPE 88.81%, R^2 0.601 을 기록하였다. 이는 전체 분산의 약 60%를 설명하면서, 예측 오차도 절반 수준으로 낮춘 결과이다. 특히 MAPE 가 90% 미만으로 낮아졌다는 점은 모델의 예측 안정성이 향상되었음을 의미한다.

XGBoost 모델은 네 가지 지표 중 세 가지에서 가장 우수한 성능을 보였다. RMSE 는 3,387.67 로 가장 낮았으며, R^2는 0.606 으로 가장 높은 설명력을 보였다. MAE 는 1,781.62, MAPE 는 87.09%로 각각 Random Forest 대비 소폭 높은 수준이지만, 전반적으로 비슷한 수준의 예측력을 보여준다. 이는 XGBoost 가 예측 정확도와 모델 안정성 측면에서 본 과제에 가장 적합한 모델로 판단되는 근거가 된다.

4.2 Feature Importance



모델별 변수 중요도 분석 결과, Random Forest와 XGBoost는 서로 다른 방식으로 변수의 중요도를 평가하고 반영하는 경향을 보였다.

먼저, Random Forest 모델은 XGBoost에 비해 다양한 변수들에 고르게 중요도를 부여하였다. 가장 높은 중요도를 보인 변수는 rolling_7(최근 7일간 평균 방문객 수)로 약 **22%**의 비중을 차지하였다. 이어서 lag_7, weekend, lag_1, day_of_week 등의 파생변수가 뒤를 이었으며, 이는 Random Forest가 시계열적 패턴 및 계절성, 반복성을 보다 세분화하여 반영하고 있음을 의미한다. 또한 기온(tmax, tmin)과 습도(humidity), 미세먼지(air) 등 기상 및 대기 환경 변수들도 일정 수준의 중요도를 보이며 예측에 기여한 것으로 나타났다.

반면, XGBoost 모델의 경우, weekend(주말 여부) 변수가 전체 중요도의 70% 이상을 차지하며 압도적으로 높은 중요도를 기록하였다. 이는 XGBoost가 예측 성능의 대부분을 주말 여부에 의존하고 있으며, 주말과 평일 간 입장객 수 차이가 모델 학습에 있어 결정적인 역할을 한다는 점을 강하게 시사한다. 그 외에는 rolling_7(7일 이동평균), day_of_week(요일), lag_7(7일 전 수요) 등의 시계열 기반 파생변수들이 상대적으로 낮은 수준의 기여도를 보였다. 기상 변수(tmax, tmin, humidity 등)나 미세먼지(air), 강수량(rain),

방학 여부(vacation) 등의 변수는 전반적으로 매우 낮은 중요도를 보였으며, snow(적설)는 거의 사용되지 않은 것으로 나타났다.
vacation, snow, year 등의 변수는 두 모델 모두에서 낮은 중요도를 기록하며 예측력 기여가 미미한 것으로 평가되었다. 결론적으로, 입장객 수 예측에서 시계열 파생변수(rolling 평균, lag 변수)의 영향력이 매우 크며, 주말 여부는 모든 변수 중 가장 핵심적인 요인으로 확인되었다. 이는 향후 운영 계획 수립 시 주말·평일 구분과 더불어, 직전 주간의 방문 추세를 기반으로 수요를 예측하는 전략이 효과적임을 시사한다.

5. 결론

본 연구에서는 서울어린이대공원의 입장객 수 예측을 위하여 공공 데이터를 통합한 시계열 데이터셋을 기반으로 다양한 예측 모델을 구축하고, 그 성능을 비교·분석하였다.

모델링 결과, Naïve 및 Seasonal Naïve 모델은 계절성과 외부 요인을 반영하지 못해 매우 낮은 예측 성능을 보였으며, R^2 역시 0 또는 음수로 측정되어 기준선으로서의 역할만을 수행하는 데 그쳤다. 반면, 머신러닝 기반의 Random Forest와 XGBoost 모델은 전반적으로 우수한 성능을 보였으며, 특히 XGBoost 모델은 가장 낮은 RMSE(3,387.67)와 가장 높은 R^2(0.606)을 기록하여 본 연구의 최적 모델로 평가되었다. Random Forest 모델도 유사한 수준의 성능을 보여, 예측의 안정성과 변수 해석 가능성 측면에서 유의미한 대안을 제공하였다.

변수 중요도 분석을 통해 입장객 수 예측에 가장 큰 영향을 미치는 요인으로는 주말 여부와 최고기온이 확인되었으며, 이는 공원 운영 측면에서도 전략적으로 활용 가능한 통찰을 제공한다. 예를 들어 주말 및 기온 상승 시 입장객 수가 급증할 가능성이 높아, 인력 배치, 시설 가동률, 마케팅 전략 등을 사전적으로 조정할 수 있는 기초 자료로 활용될 수 있다.

향후 연구에서는 시간대별 입장 패턴, 인근 지역 이벤트, 대중교통 이용량 등 추가적인 외부 요인을 반영하거나, 딥러닝 기반의 시계열 예측 모델(LSTM 등)을 적용함으로써 예측 정밀도를 더욱 향상시킬 수 있을 것으로 기대된다. 또한 SHAP 기반의 변수 중요도 해석 기법을 추가적으로 활용할 경우, 모델의 투명성과 신뢰도를 높이는 데에도 기여할 수 있을 것이다.