디지털 트윈 기반의 이종 데이터를 활용한 공공 자전거 수요량 예측

김소현 [1, 2], 백명선 [1, 3]
[1] 한국전자통신연구원, [2] 가천대학교 컴퓨터공학과, [3] 과학기술연합 대학원대학교
ts7704@gachon.ac.kr, sabman@etri.re.rk

Prediction of Public Bicycle Demand Using Digital Twin-Based Heterogeneous Data

So-Hyeon Kim[1, 2], Myung-Sun Baek[1, 3]
[1]ETRI, [2]Gachon University, [3]UST

요 약

본 논문에서는 과거의 기상관측 데이터를 기반하여 공공자전거 수요량을 예측해 다음날 공공자전거 배치 수에 도움이 되고자 실험을 진행하였다. 디지털 트윈 기반의 이종 데이터를 활용한 공공 자전거 수요량 예측 모델로 앙상블의 부스팅 기법인 XGBoost 을 제시하고, 실제 방재기상관측 데이터세트와 서울시 공공자전거 이용현황 데이터세트로 다양한 전처리 과정을 수행하여 실험한 결과를 도출하였다.

Ⅰ. 서 론

디지털 트윈 기술은 현실 세계의 객체 (사람, 사물, 공간)를 디지털로 복제하고, 다양한 모의실험을 수행하여, 현실 세계의 문제들을 해결하거나 최적화하기 위해 개발된 기술이다. 객체를 디지털로 전환하거나, 디지털 전환된 객체를 통한 예측적 모의실험을 수행하기 위해서는 Advanced machine learning, 데이터 분석, 데이터 및 객체 시각화, Advanced simulation 등 최신의 기술들이 요구된다 [1].

본 논문에서는 기상 디지털 트윈, 공공자전거 관리 디지털 트윈을 가정하여, 해당 디지털 트윈 출력에 해당하는 기상관측 데이터와 공공자전거 수요량 데이터를 활용하여 향후 공공자전거 수요량을 예측 및 관리할 수 있는 기술을 연구하였다.

Ⅱ. 이종 데이터 및 전처리

본 논문에서는 2017년 01월 01일부터 2017년 12월 31 일까지 시계열 데이터(서울시 일별 방재기상관측 데이터세트, 일별 공공자전거 이용 현황 데이터세트)를 선택하여 두 데이터의 특성 및 상관관계 등을 분석하였다.

일별 방재기상관측 데이터세트의 변수는 평균 기온, 최고 기온, 최저 기온, 일 강수량, 순간 최대 풍속, 일교차로 이루어져 있다. 그리고 방재기상관측 데이터에 기반하여 공공자전거 수요량 예측의 성능을 높이기 위해, 일별 방재기상관측 데이터세트에 일교차 변수 칼럼을 추가했다. 일교차 변수는 (최고기온 – 최저기온)으로 계산한 값이다. 표 1 은 방재기상관측 데이터의 변수와 공공자전거 수요량 변화와의 상관관계의 예측을 나타낸다.

표 1. 방재기상관측 변수의 수요량에 대한 예상 효과

방재기상관측 변수	변수의 예상 효과
평균, 최고, 최저 기온	기온이 매우 높거나 낮으면 수요량 감소
일 강수량	강수량이 많은 여름에 수요량 감소
순간 최대 풍속	풍속이 강할 때 수요량 감소, 약할 때 증가
일교차	일교차 큰 경우 수요량 감소, 작은 경우 증가

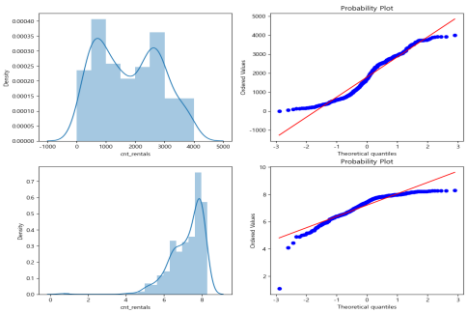

그림 1. 자전거 수요량 데이터 로그 스케일링 전/후

데이터 전처리를 위해 본 논문에서 Feature Engineering 방식을 사용하였다. Feature Engineering 은 데이터를 기반으로 한 예측 성능을 향상시키기 위해, 데이터를 Feature 로 변환하고 가공하는 프로세스이다 [2]. 본 프로세스를 통해 이상치 제거, 데이터 스케일링 등을 시각화 하여 수행하였다.

(1) 자전거 수요량 데이터 로그 스케일링(Log scaling)

방재기상관측 데이터세트와 자전거 수요량 데이터세트 사이의 수치 값 차이가 매우 크다 (방재기상관측 데이터 최댓값: 150, 자전거 수요량데이터 최댓값: 1791). 데이터의 분포가 서로 크게 다르면, 모델은 데이터 분포가 좁은 특성 값의 변화는 결과 값을 예측하는데 영향이

적다고 판단하여 모델 성능이 낮고 오차가 크게 나올 것이다[3]. 따라서 그림 1 과 같이 자전거 수요량 데이터에 로그 스케일링을 적용하였다.

(2) 일 강수량이 0.0 인 데이터 처리

방재기상관측 데이터세트의 일 강수량 데이터는 우리나라 기후 특성상 여름에 비가 많이 오기 때문에 그림 1 과 같이 계절별 차이가 크고 대부분의 값이 0.0 이다. 따라서 강수량 데이터의 의미를 보강하기 위해 일 강수량이 0.0 인 값에 Random Forest 모델을 적용하였다[5]. 이 절차를 통해 강수량이 0.0 인 값을 Random Forest 로 예측된 값으로 대체되었다.

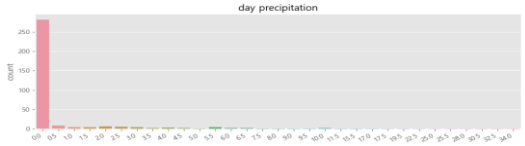

그림 2. 일 강수량 데이터 히스토그램

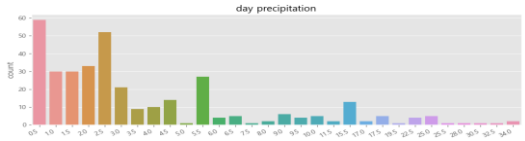

그림 3. 일 강수량 데이터 예측한 후 히스토그램

(3) 이상치 제거(Outlier removal)

이상치는 일반적인 데이터 패턴과 다른 패턴을 가지고 있는 데이터를 뜻한다. 이상치를 포함하여 예측할 경우 왜곡된 예측 결과를 얻게 되므로 전처리 과정에서 이상치를 확인하고 제거하는 과정은 필수이다 [4]. 본 연구에서는 기존 데이터의 일 강수량, 순간 최대 풍속에 존재하는 이상치 데이터를 제거하여 데이터를 보완하였다.

Ⅲ. 예측 모델 및 방법

본 연구에서는 II 장에서 설명된 전처리 과정을 거친 데이터를 활용하여 자전거 수요량을 예측하는 XGBoost 모델을 설계하였다. XGBoost 는 ML 방법 중 매우 효과적이고 널리 사용되는 방법 중 하나이며, 여러 개의 트리를 임의적으로 학습하여 에러를 줄이는 과정을 앙상블의 부스팅 기법이다. 병렬 학습 방법을 가져 약한 분류기를 강한 분류기로 만든다 [6].

그림 4 에서와 같이 전처리가 완료된 데이터를 입력 값으로 이용하여 해당 일의 공공자전거 수요량을 예측하고 실제 값과의 차이를 비교하여 성능을 검증한다.

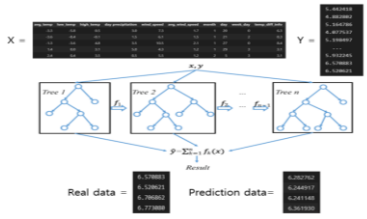

그림 4. XGBoost 모델 기반 입력/출력

Ⅳ. 실험 및 결론

논문에서 개발된 기술의 성능 검증을 위해 2017 년 방재기상관측 데이터와 공공자전거 이용 현황 데이터를 활용하였다. 총 365 일에 해당하는 365 행 중 이상치를 제거하고 303 행을 활용하였으며, 이중 20%를 테스트 데이터, 나머지를 학습 데이터로 사용하였다. XGBoost 모델을 사용하여 기상관측 데이터 기반으로 공공자전거 수요량 예측을 하였을 때 공공자전거 수요량과 관련성이 높은 변수는 그림 5 와 같이 일교차, 평균 기온, 최고 기온, 일 강수량 순이다.

하이퍼파라미터는 Optuna 프레임워크를 사용하여 최적화하였다. 성능 평가 지표는 NMAE 를 사용하였다. XGBoost 모델을 사용하여 공공자전거 수요량을 예측한 결과 NMAE 0.064 로 오차가 약 6%이며 그림 6 은 실제 데이터와 예측 데이터 결과를 보여준다.

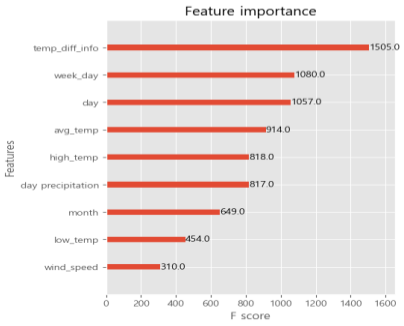

그림 5. 속성(변수) 중요도

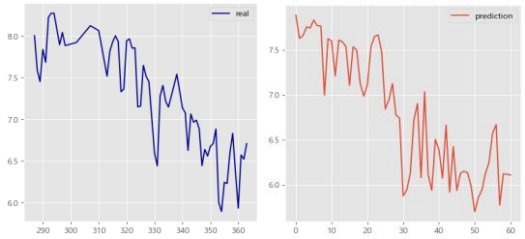

그림 6. 공공 자전거 수요량 예측 성능

(왼쪽: 실제 값, 오른쪽: 예측 값)

데이터 전처리 과정 중 Feature Engineering 의 사용 여부 차이는 실제 값과 예측 값의 오차에서 많은 차이를 보였다. Feature Engineering 수행 전에는 오차가 약 30%으며, 로그 스케일링, 이상치 제거 등의 전처리 후 오차가 약 6%로 감소하였다. 본 논문에서 제시한 기술을 활용하면 기상 데이터를 기반으로 공공자전거 수요량을 효과적으로 예측하는 것이 가능하며, 이를 기반으로 공공자전거 관리에 효율을 높일 수 있다.

ACKNOWLEDGMENT

이 논문은 2023 년도 정부 (과학기술정보통신부) 의 재원으로 정보통신기획평가원의 100% 지원을 받아 수행된 연구임 (No. 2022-0-00545, (2 세부) 지능형 디지털 트윈 연합 객체 구성 및 데이터 프로세싱 기술 개발)

참 고 문 헌

[1] 백명선, 정득영, 이용태. ” 연합 디지털 트윈을 위한 데이터 유효성 검증 방법 연구,” 한국통신학회 학술대회논문집, pp. 421-422.

[2] Sinan Ozdemir, Divya Susarla. “Feature Engineering Made Easy, Packt,” 2018.

[3] 김영원, 이수진. “XGBoost 기반 침입탐지모델을 위한 데이터 스케일링 및 특성선택 기법 연구,” 한국컴퓨터정보학회 학술발표논문집, pp. 251-254.

[4] 홍예진, 나은희, 정용환, 김양우. “대용량 데이터 분석을 위한 이상치 제거용 분산처리 환경,” 한국컴퓨터정보학회 학술발표논문집, pp. 73-74.

[5] 문재욱, 정승원, 김형준, 황인준. “랜덤 포레스트를 활용한 지역별 하루 단위 감염병 발생 예측,” 한국정보과학회 학술발표논문집, pp. 335-337.

[6] Tianqi Chen and Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.