JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 6, pp. 1305-1315, Jun. 2023

Check for updates

공공 자전거 수요 예측을 위한 사이킷런의 지도 기계 학습 모델 성능 비교

권 혜 진[1] · 하 진 영[2*]
[1]강원대학교 컴퓨터정보통신공학과 석사
[2*]강원대학교 컴퓨터공학과 교수

Performance Comparison of Scikit-Learn's Supervised Machine Learning Models for Public Bicycle Demand Prediction

Hye-Jin Kwon[1] · Jin-Young Ha[2*]

[1]M.S., Department of Computer and Communications Engineering, Kangwon National University, Chuncheon, Gangwon 24341, Korea
[2*]Professor, Department of Computer Science and Engineering, Kangwon National University, Chuncheon, Gangwon 24341, Korea

[요 약]

본 연구는 공공 자전거 수요 예측을 위해 사이킷런에서 제공하는 기계 학습 모델들의 성능을 비교 평가를 하였다. 공공에서 제공하고 있는 신뢰성 있는 데이터를 실험에 사용하였는데, 서울시가 제공하는 '서울시 공공자전거 이용정보' 데이터와 기상청이 제공하는 '날씨 정보'를 활용하였다. 사이킷런 지도학습의 모델인 랜덤 포레스트, 그래디언트 부스팅, 결정 트리, 선형 회귀를 사용하였고, 성능을 비교 분석하기 위해 RMSE, R^2, RMSLE, 정확도를 계산하여 평가 지표로 사용하였다. 그 결과 랜덤 포레스트 모델이 RMSE 347.37, R^2 0.74, RMSLE 0.51, 정확도 67.61%로 가장 성능을 보였다. 그래디언트 부스팅과 결정 트리 모델은 랜덤 포레스트보다 다소 낮은 성능을 보였지만, 선형 회귀의 성능은 현저하게 낮음을 확인할 수 있었다. 다양한 모델들을 활용한 수요 예측 분석을 통해 최적의 모델을 선정하여 수요 예측 오차를 줄여 나가는 데 도움이 될 수 있을 것으로 판단한다.

[Abstract]

This study compares and evaluates the performance of machine learning models provided by scikit-learn for predicting public bicycle demand. Reliable data provided by the government, namely "Seoul public bicycle usage information" provided by the Seoul Metropolitan Government and "weather information" provided by the Korea Meteorological Administration, were used for the experiment. Supervised learning models in scikit-learn, namely random forest, gradient boosting, decision tree, and linear regression, were used, and performance was evaluated using RMSE, R^2, RMSLE, and accuracy. The random forest model showed the best performance with an RMSE of 347.37, R^2 of 0.74, RMSLE of 0.51, and accuracy of 67.61%. The gradient boosting and decision tree were the next best-performing models, whereas the linear regression had the worst performance, as expected. Thus, from the various models for demand prediction analysis, the optimal model can be selected to reduce demand prediction errors.

색인어 : 기계 학습, 데이터 분석, 수요 예측, 공공자전거, 파이썬, 사이킷런
Keyword : Machine Learning, Data Analysis, Demand Prediction, Public Bicycles, Python, Scikit-Learn

http://dx.doi.org/10.9728/dcs.2023.24.6.1305

Received 13 April 2023; **Revised** 22 May 2023
Accepted 30 May 2023

***Corresponding Author; Jin-Young Ha**

Tel: +82-33-250-6386
E-mail: jyha@kangwon.ac.kr

Ⅰ. 서 론

‘따릉이’라는 명칭으로 2015년 10월부터 서울시에서 운영하는 공공자전거 대여 서비스는 시민들에게 편리한 근거리 이동 수단을 제공함으로써 건강하고 깨끗한 자전거 도시를 만드는 것은 물론 녹색 성장 선도 도시로서의 서울의 이미지를 제고하고 있다. 2021년 기준 따릉이 대여소는 서울에 2천 6백 개소이며 대여할 수 있는 자전거 수는 4만 대에 육박한다. 2020년 한 해 동안 대여한 자전거는 2,370만건에 달할 정도로 인기가 있다. 이는 이전 연도에 비해 24% 증가한 수치이며 고유가 시대에는 특히 수요 증가폭이 확대될 것으로 예상된다[1]. 그림 1은 서울 공공자전거 연도별 이용 현황을 보여주고 있다[2]. 그런데, 아무리 자전거를 확보하더라도 수요 예측에 실패한다면 어떤 대여소는 대여할 자전거가 남아돌고, 다른 대여소는 대여할 자전거가 없어 시민들이 불편을 겪을 수 있다. 이 문제를 해결하기 위해 기계 학습을 이용한 수요량 예측을 함으로써 각 대여소에 적절한 대수의 자전거를 사전에 배치하는 데 도움을 줄 수 있는 연구를 수행하였다.

예측은 행정, 의료, 문화, 비즈니스 전반에 이르기까지 더 현명한 의사 결정을 할 수 있는 자원을 제공해주며 업무의 효율성을 가져다줄 수 있다. 예측 분석은 주로 과거의 데이터를 분석하여 미래를 예측하는 방식인데 기계 학습, 통계 모델링, 데이터 마이닝 같은 분석 기술을 사용한다[3]-[6]. 이 중 기계 학습은 데이터를 기반으로 패턴을 학습하고 결과를 예측하는 알고리즘 기법이다. 최근 몇 년간의 기계 학습과 딥러닝의 발전은 다양한 분야에 예측 모델을 사용할 기회를 만들어주었다. 데이터 분석 영역은 기계 학습 알고리즘 기반으로 더욱 정확한 예측과 의사 결정을 도출하고 있다. 기계 학습 알고리즘은 데이터를 기반으로 통계적인 신뢰도를 강화하고 예측 오차를 최소화하기 위한 다양한 수학적 기법을 적용해 데이터 내의 패턴을 스스로 인지하고 신뢰도 있는 예측 결과를 도출한다. 데이터를 관통하는 패턴을 학습하고 이에 기반한 예측을 수행하면서 데이터 분석 영역에 새로운 혁신을 가져오고 있다.

최근 기계 학습의 발전은 통계나 수학 이론보다 경험을 바탕으로 발전하는 경우도 많다. 컴퓨터 과학 분야는 이런 발전을 주도하고 있는데, 대표적인 기계 학습 라이브러리는 사이킷런(scikit-learn)이다[7]-[9]. 사이킷런 라이브러리는 파이썬 API를 사용하는 데 기계 학습의 관심이 높아지면서 파이썬과 함께 사이킷런 라이브러리가 큰 인기를 얻고 있다. 사이킷런에는 모든 기계 학습 알고리즘이 포함되어 있지는 않다. 많은 사람이 검증하고 사용한 다음 장단점을 파악하게 되면서 이런 알고리즘이 유익하다고 증명되면 사이킷런 라이브러리 개발자들은 이 알고리즘을 라이브러리에 추가한다. 그러므로 기계 학습 라이브러리에 포함된 알고리즘들은 안정적이며 성능이 검증되어 있다고 볼 수 있다. 사이킷런과 같은 오픈 소스 라이브러리의 발전 덕분에 기계 학습 분야는 말 그대로 폭발적으로 성장하였다[3],[7]-[9].

그림 1. 서울시 공공자전거 연도별 현황
Fig. 1. Seoul public bicycles by year

기계 학습 기반의 수요 예측 분석이 본격적으로 진행된다면 이제 무엇보다 중요한 것은 신뢰성 있는 다량의 데이터 확보다. 데이터를 이해하고 효율적으로 가공, 처리, 추출해 최적의 데이터를 준비하는 것이 필요하다. 이를 기반으로 알고리즘을 구동할 수 있도록 준비하는 것이 데이터 분석에 있어서 가장 중요한 단계라고 볼 수 있다. 기계 학습, 특히 딥러닝 알고리즘을 누구나 쉽게 사용할 수 있는 상황에서 경쟁력은 어떠한 품질의 데이터로 만든 기계 학습 모델이냐에 따라 결정될 수 있다. 또한, 데이터의 신뢰성이 떨어진다면 성능 좋은 알고리즘이 존재한다고 하더라도 정확한 예측은 어려울 것이다.

이를 위해 신뢰성 있는 데이터의 확보가 중요하다. 정부는 국가가 보유하고 있는 다양한 데이터를 『공공데이터의 제공 및 이용 활성화에 관한 법률(제11956호) 법제처 국가법령정보센터(https://www.law.go.kr/법령/공공데이터의제공및이용활성화에관한법률)』(약칭: 공공데이터법)에 따라 국민이 더욱 쉽게 공유하고 활용할 수 있도록 공공데이터를 제공하고 있다.

본 연구에서는 서울시의 ‘따릉이’ 공공자전거 대여량을 분석 대상으로 하였다. 자전거 대여 수요는 날씨가 많은 영향을 줄 수 있으므로 날씨 정보도 모델 학습에 사용하였다. 실험에 필요한 데이터 확보를 위해 서울시가 ‘서울 열린데이터광장’[10]을 통해 제공하고 있는 공공데이터 중 ‘서울시 공공자전거 이용정보’ 데이터와 기상청이 ‘기상자료개방포털’[11]을 통해 제공하고 있는 날씨 정보를 활용하였다. 본 연구에서는 사이킷런의 기계 학습 모델 중 지도학습 기반의 예측 모델을 통해 대여량을 예측하고 각 모델의 성능을 분석하고자 한다. 이러한 수요 예측 결과는 공공 자전거 관리 및 업무 효율성에 도움을 줄 것으로 본다.

Ⅱ. 연구배경

2-1 선행연구

국내 학술지와 학위 논문을 검색할 수 있는 DBpia[12] 기준으로 ‘공공자전거’에 대한 검색 결과는 378건인 것에 비해 ‘공공자전거 예측’에 대한 검색 결과 46건에 불과하다. 공공자전거에 관한 관심에 비해 예측에 관한 연구 사례는 드물다

고 볼 수 있다. 수요 예측에 연구가 발표되는 학회를 기준으로 봤을 때는 주로 교통학회와 정보과학회로 나누어진다. 교통과 정보기술이라는 두 축으로 수요 예측이 연구되고 있지만 최적의 방안을 탐색하기 위한 연구라는 측면에서는 공통점을 가지고 있다.

기계 학습 알고리즘을 적용하여 대여량을 예측한 연구를 살펴보면 '랜덤 포레스트를 이용한 대전시 공공자전거 수요 예측 (2017)'[13] 연구가 있다. 자전거 대여 이력과 기상정보, 축제 데이터를 수집하여 시민들의 자전거 이용 패턴을 파악하고 랜덤 포레스트(random forest) 알고리즘으로 예측 모델을 구현하였으며 RMSE(root-mean-square error) 방법을 이용하여 평균 오차율을 구하였다. 평균 오차율은 0.015592로 나타났다.

'딥러닝 모형을 활용한 공공자전거 대여량 예측에 관한 연구(2020)'[14]는 딥러닝 모형을 개발하여 평가하는 것으로 기상 자료, 지하철 이용량 자료를 수집하여 데이터셋을 구성하였고 지수 평활 모형(exponential smoothing model), ARIMA(AutoRegressive Integrated Moving Average) 모형, LSTM(long short-term memory) 기반의 딥러닝 모형을 구축한 후 MSE(mean squared error)와 MAE(mean absolute error) 두 가지 평가 지표를 사용하여 예측 오차를 비교하였다. 지수 평활 모형으로 MSE 348.74, MAE 14.15 값이 산출되었다. ARIMA 모형으로 MSE 170.10, MAE 9.30 값을 얻었으며 딥러닝 모형으로 MSE 120.22, MAE 6.76 값이 산출되었다.

'서울시 공공자전거 수요예측 모형 비교 연구(2021)'[15]는 서울시 공공자전거 대여 이력 데이터를 분석 대상으로 하였고, 자전거 수요 예측을 위해 다양한 기법들을 모델링하여 비교하였다. 딥러닝 기반의 LSTM, 서포트 벡터 회귀 모형(SVR:support vector regression), 군집별 수요 예측 모형(VAR:vector auto regression)을 대상으로 하였고, 예측모형의 평가는 RMSE와 MAE를 사용하였다. LSTM(ReLU)의 경우 RMSE 6.08, MAE 4.07, SVR의 경우 RMSE 7.072 MAE 5.23, VAR의 경우 RMSE 205.63, MAE 97.47의 결과가 도출되었고, LSTM, SVR, VAR 모형 순으로 예측력이 좋게 나왔다.

이러한 연구 외에도 구글의 예측 모델 분석 대회 플랫폼 캐글(kaggle) (bike sharing demand, 2016)과 국내의 인공지능 경진대회 플랫폼 데이콘 (서울시 따릉이 대여량 예측 경진대회, 상시) 등을 통해 공공자전거 대여량 수요를 예측하기 위한 프로젝트들이 진행되고 있다.

2-2 선행연구의 검토

예측 모델을 선정하는 데 있어 다양한 기법들을 사용하여 결과를 도출하는 것은 필요한 과정이다. 이러한 연구들은 동일한 데이터에 대해 다양한 기계 학습 모델을 사용하여 결과를 비교하고 최적의 모델을 선정하는 데 도움을 준다. 실제로 다양한 분야에서 최적의 기계 학습 모델을 선정하기 위한 성능 비교 연구는 계속되고 있다. DBpia 기준으로 '기계 학습 예측 비교'에 대한 검색 결과는 908건으로 국방, 환경, 교육, 경제, 산업 전반에 기계 학습을 활용한 예측 모델 비교 연구가 이루어지고 있다.

공공자전거 대여량 수요 예측의 경우, 공공자전거 도입 초기에는 특정 모델을 사용하여 대여량을 예측한 연구가 진행되었다면, 최근에는 최적의 모델을 선정하기 위한 다양한 비교 연구와 프로젝트들이 진행되고 있다. 다만, 수요 예측과 관련하여 딥러닝 기반의 연구만이 있어 다수의 데이터 분석가들이 사용하고 있는 기계 학습 알고리즘에 대한 공공자전거 수요 예측 성능 비교 분석 연구가 필요한 상황이다.

Ⅲ. 데이터 분석

3-1 분석방안

기계 학습에서 주로 사용되는 프로그래밍 언어로 주로 R과 파이썬(Python)이 많이 사용되는데 R은 전통적인 통계 전용 프로그램 언어이며 파이썬은 다양한 영역에서 사용되는 개발 전문 프로그램 언어인데, 최근 인공지능 라이브러리 호출에 많이 사용되고 있다. 그림 2는 최근 5년간의 파이썬과 R에 대한 관심을 구글트렌드를 통해 살펴본 결과이다. Python Machine Learning, Python Data Science, R Machine Learning, R Data Science의 키워드로 R과 파이썬의 관심도 변화를 비교하여 보았다. 구글트렌드의 추이를 살펴보면 R 기계 학습보다 파이썬 기계 학습이 더 많은 관심을 받았다고 볼 수 있으며 파이썬 기계 학습에 대한 관심은 2022년에 큰 폭으로 증가하였음을 볼 수 있다.

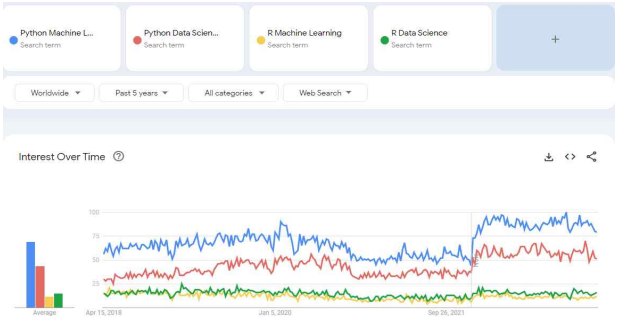


그림 2. 파이썬 기계 학습의 구글트렌드 추이
Fig. 2. Google Trends in Python machine learning

그림 3은 데이터 분석 대회가 열리는 온라인 플랫폼인 캐글의 설문조사 중 캐글 이용자들이 데이터를 이용하고 분석되는데 가장 많이 사용하는 기계 학습 알고리즘에 대한 조사 결과이다. 설문 조사 결과 가장 많이 사용하는 기계 학습 알고리즘은 선형/로지스틱 회귀(linear/logistic regression)이

다. 데이터 분석가, 데이터 과학자, 기계 학습 엔지니어, 소프트웨어 엔지니어 직군 모두 선형/로지스틱 회귀를 사용한다고 답한 사람이 가장 많았다. 기계 학습 알고리즘에서 두 번째로 많이 사용되는 것은 결정 트리(decision tree)와 랜덤 포레스트이며, 그래디언트 부스팅(gradient boosting)의 인기가 높아짐에도 여전히 선형 회귀(linear regression), 결정 트리, 랜덤 포레스트의 사용 추이는 꾸준하게 상위를 유지하고 있다[16].

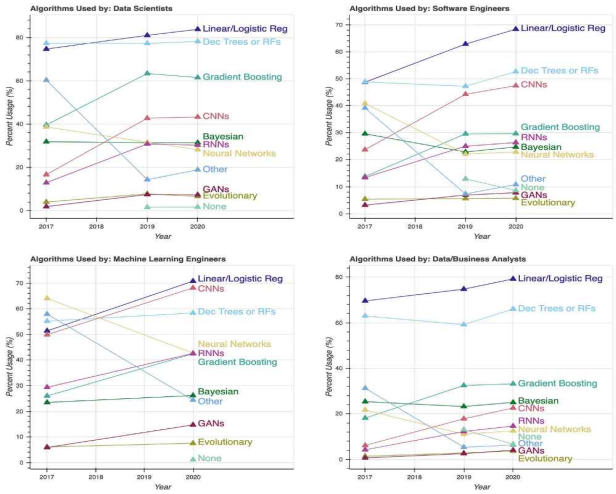


그림 3. 캐글 이용자가 사용하는 기계 학습 알고리즘 추이
Fig. 3. Trends in machine learning algorithms used by Kaggle users

다양한 분야에서 가장 많이 사용되는 알고리즘인 선형 회귀는 종속 변수와 한 개 이상의 독립변수 간의 선형 상관 관계를 모델링하는 회귀분석 기법이다. 로지스틱 회귀(logistic regression)의 경우에는 0 또는 1과 같은 이진 결과를 예측하는 데 사용되는 기법이므로 연속 변수인 자전거 수요 예측에는 적합하지 않아 본 연구에서는 다루지 않는다.

결정 트리는 데이터 마이닝에서 일반적으로 사용되는 방법론으로, 귀납추론을 위해 자주 사용되는 실용적인 방법이다. 이것은 데이터들을 트리 구조의 루트에서 시작하여 차례로 중간 노드들을 거쳐 단말 노드에 배정하는 기능을 수행한다. 결정 트리의 모델은 스무고개와 같이 데이터의 패턴을 잘 나눌 수 있는 질문을 계속 추가하면서 분류 정확도를 높일 수 있다. 그 모양이 '나무'와 같다고 해서 결정 트리라고 부른다[17].

랜덤 포레스트는 분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종이다. 앙상블은 여러 기계 학습 모델을 연결하여 더 강력한 모델을 만드는 기법이다. 다수의 결정 트리로부터 그 결과를 평균냄으로써 과적합을 줄일 수 있으며 검증 세트와 테스트 세트에 안정적인 성능을 얻을 수 있다[18]. 그래디언트 부스팅은 회귀 및 분류 작업에 사용되는 기계 학습 기술이다. 깊이가 얕은 결정 트리를 사용하여 이전 트리의 오차를 보완하는 방식으로 앙상블 하는 방법이다[19].

본 연구에서는 파이썬 기계 학습 기법을 사용하여 예측 분석을 실시하고자 한다. 파이썬 기계 학습을 라이브러리 중에서 가장 사용하기 편리한 개발 라이브러리인 사이킷런을 사용하고[19], 사이킷런에서 제공하는 지도학습의 모델인 랜덤 포레스트, 그래디언트 부스팅, 결정 트리, 선형 회귀 모델을 사용하여 공공자전거 대여량 수요 예측을 실시하고 각각의 결과를 분석 비교하도록 하겠다.

3-2 데이터 개요 및 최종 데이터 셋 구성

공공자전거 대여량 예측을 위해 서울시에서 제공하고 있는 공공데이터를 활용하였고 대여량에 변수로 작용할 수 있는 날씨의 영향도를 분석하고자 기상청에서 제공하는 날씨 정보를 포함하여 분석하였다. 표 1은 분석 대상인 '서울시 공공자전거 이용정보' 데이터와 기상청에서 제공하는 '날씨 정보' 데이터의 칼럼과 칼럼의 타입을 보여준다. 이 두 개의 데이터셋에서 공공자전거 이용정보 '대여시간'과 날씨 정보의 '일시' 정보를 매핑하였다. 대여량 예측에 특징으로 사용할 시간대별 이용건수와 날씨 정보를 구성하는 기온, 강수량, 풍속, 습도 정보를 표 2와 같이 최종데이터 셋으로 구성하였다. 공공자전거의 이용정보 데이터의 경우 반기별로 데이터를 제공하므로 최종 업로드된 자료인 2022년 6월 30일까지 자료를 기준으로 최근 2년 치 데이터(5.13G)를 분석하였고 날씨 정보도 2020년 7월 1일부터 2022년 6월 30일까지 2년 치 데이터를 같이 포함하였다. 전체 데이터 중 대여량 수요 예측에 사용할 최종 데이터셋 구성을 위해 전체 데이터의 20%를 비복원 추출하여 17,435건의 최종 데이터 셋을 만들었다.

2020년 7월부터 2022년 6월까지 대여량 추이를 알기 위해 데이터 시각화 작업을 진행하였다. 데이터 시각화 작업은 파이썬의 matplotlib, seaborn 라이브러리를 사용하였다. 그림 4는 2020년 7월부터 월별 대여량 추세를 보여주고 있다. 2021년 6월부터 시작되는 그래프와 비교했을 때 전반적으로 대여량이 적음을 볼 수 있으며 특별히 2020년 8월은 코로나 2차 대유행으로 사회적 거리 두기 2단계와 2.5단계가 발생한 시기로 대여량이 급감하였음을 볼 수 있다.

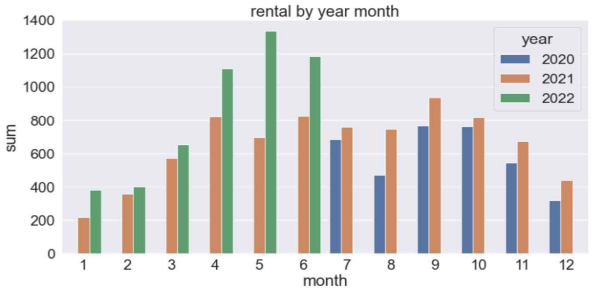
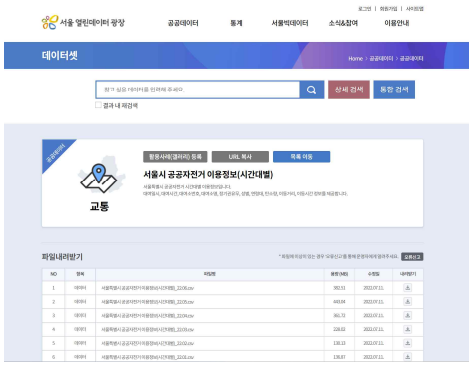


그림 4. 월별 대여량 추세
Fig. 4. Rental by year month

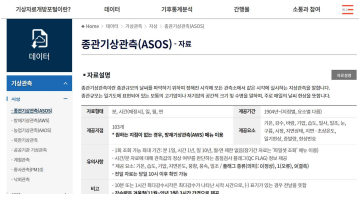
표 1. 원본 데이터
Table 1. Source data

<Public bicycle usage>



No	Column Name	Type
1	rental hours	int64
2	STATION_ID	int64
3	station name	object
4	rental classification code	object
5	gender	object
6	age code	object
7	number of rentals	int64
8	exercise amount	object
9	carbon amount	object
10	moving distance	float64
11	rental time(hour)	float64
12	moving dist(M)	float64
13	rental time(min)	float64

<Weather info>



No	Column Name	Type
1	place	int64
2	place name	object
3	date	datetime64
4	temp(°C)	float64
5	precipitation(mm)	float64
6	wind speed(m/s)	float64
7	humidity(%)	int64

표 2. 최종 데이터셋
Table 2. Final dataset

year	month	date	hour	amount	time	temp	precipitation	wind speed	humidity	day of week
2020	7	1	1	255	2020-07-01 1:00	18.8	0	3	87	2
2020	7	1	2	170	2020-07-01 2:00	18.7	0	3	87	2
2020	7	1	3	190	2020-07-01 3:00	18.4	0	2.3	86	2
…	…	…	…	…	…	…	…	…	…	…
2022	6	30	21	123	2022-06-30 21:00	21.4	1	1.9	99	3
2022	6	30	22	288	2022-06-30 22:00	21.5	1	1.5	99	3
2022	6	30	23	370	2022-06-30 23:00	21.7	0.1	0.8	99	3

3-3 데이터 분석 및 시각화

2021년 6월 이후의 그래프 역시 2022년 4, 5월 대여량이 급증하였는데 2022년 4월 거리두기 해제로 자전거 대여량이 급속도로 증가하였다. 전체적인 월별 대여량 추세를 볼 때 하반기 데이터를 적용하면 7, 8월 이후의 대여량도 증가할 것으로 보인다. 이러한 데이터의 특이점들은 예측 데이터에 사회적인 이슈가 변수로 작용한다는 것을 보여주는 사례이다. 상대적으로 이 이슈에 영향이 없는 경우에는 전반적으로 1월부터 대여량이 증가하여 9월 정점을 이루고 하향되는 형태로 대여량이 분포되어 있음을 볼 수 있다.

그림 5는 월별 대여량의 평균 그래프로 2년간의 월별 대여량의 월별 평균값을 보여준다. 그림 6은 시간별 대여량을 그래프로 나타낸 것인데, 오전 5시를 기점으로 점점 대여량이 많아지다가 오전 8시 출근 시간에 대여량이 급증하며 18시 퇴근 시간을 기점으로 점차 줄어드는 것을 확인할 수 있다. 그림 7은 요일별 대여량으로 주중인 월요일부터 금요일까지 대여량이 주말에 비해 조금 더 많은 것으로 보인다. 주중은 월요일이 가장 적고 금요일이 가장 많은 것을 볼 수 있다.

박스 플롯(box plot)을 사용하여 데이터를 시각화하여 보았다. 박스 플롯은 사분위 수를 사용하여 데이터를 시각화하는 방법이다. 박스 플롯은 자료로부터 얻어낸 통계량인 5가지 요약 수치를 가지고 그리는데, 5가지 요약 수치란 최솟값, 제1사분위(Q1-전체 데이터 중 하위 25%), 제2사분위(Q2-데이터의 중앙값), 제3사분위(Q3-전체 데이터 중 상위 25%), 최댓값을 일컫는 말이다. 최솟값과 최댓값을 넘어가는 위치에 있는 값을 이상치(outlier)라고 부른다. 전반적인 모습은 바 플롯(bar plot)과 비슷하게 분석할 수 있지만 박스 플롯은 데이터의 분포와 이상치를 보여주고 밀집도를 동시에 보여주기 때문에 서로 다른 데이터군을 쉽게 비교할 수 있다.

그림 5. 월별 평균 대여량 추세
Fig. 5. Rental by month mean

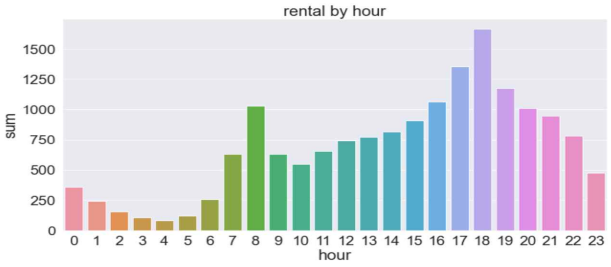


그림 6. 시간대별 대여량 추세
Fig. 6. Rental by hour

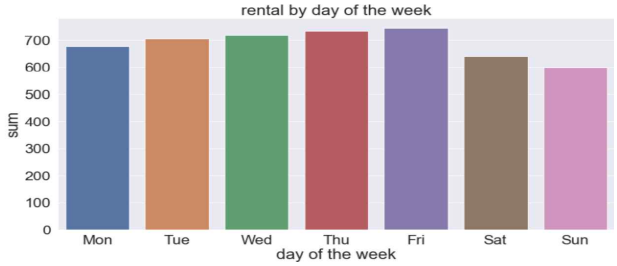


그림 7. 요일별 대여량 추세
Fig. 7. Rental by week

그림 8의 박스 플롯을 통해 출근 시간대인 8시와 퇴근 시간대의 18시에 밀집도가 크다는 것을 확인할 수 있다. 그림 9에서는 수요일과 목요일에 이상치가 있음을 발견할 수 있다. 그림 10의 경우 4, 5, 6월에 밀집도가 크다는 것을 확인할 수 있으며, 그림 11을 통해 2020년 8월에 밀집도가 급감하였음을 볼 수 있다. 전반적으로 봤을 때 여름에는 점차 밀집도가 증가하고 겨울에는 점차 밀집도가 감소하며, 2022년 4월부터는 밀집도가 크게 증가하였음을 확인할 수 있다.

시간대별 대여량 분석을 위해 포인트 플롯을 사용하여 그래프를 그려보았다. 그림 12의 첫 번째 그래프를 통해 출퇴근 시간이 대여량이 많은 것을 확인할 수 있고 두 번째 그래프에서는 주중은 출퇴근 시간대에 영향을 받지만, 주말은 주중에 비해 완만한 형태를 이룬다는 것이 시각적으로 표현된다. 주말에는 점심 시간대부터 오후까지 대여율이 증가하다가 17시를 기점으로 점차 줄어드는 것을 확인할 수 있다. 세 번째 그래프는 시간대별로 월간 대여량과의 관계를 시각화한 것인데 붉은색 계열로 표현되는 겨울에는 대여량이 적고 초록색 계열로 표현되는 여름에는 대여량이 많음을 시각적으로 확인할 수 있다.

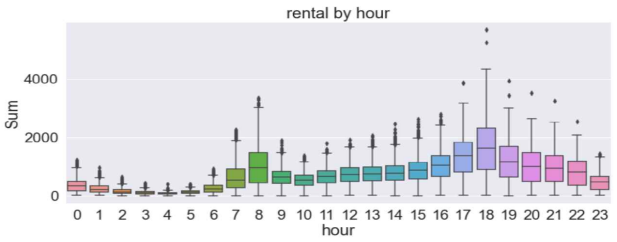


그림 8. 시간대별 대여량
Fig. 8. Rental by hour

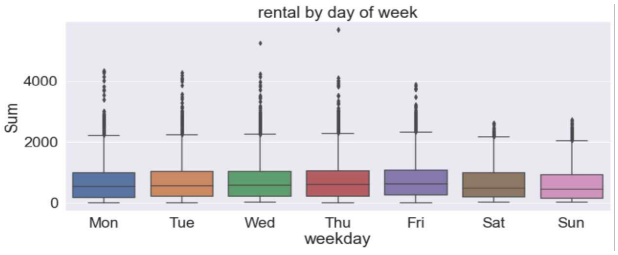


그림 9. 요일별 대여량
Fig. 9. Rental by day of week

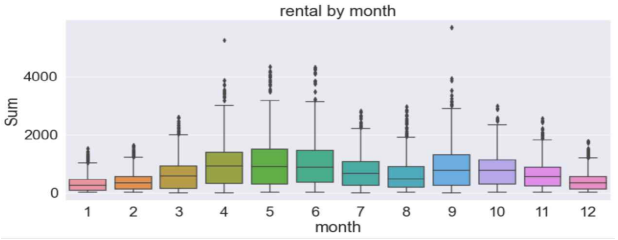


그림 10. 월별 평균 대여량
Fig. 10. Rental by month mean

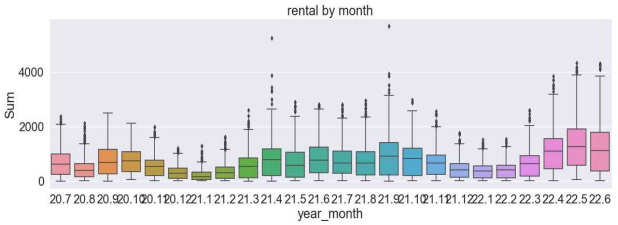


그림 11. 월별 대여량 추세
Fig. 11. Rental by month

그림 13은 시간, 온도, 강수량, 풍속, 습도, 대여량의 상관관계를 시각적으로 확인하기 위한 히트맵(heat map)이다. 히트 맵은 열을 뜻하는 히트(heat)와 지도를 뜻하는 맵(map)을 결합시킨 단어로, 색상으로 표현할 수 있는 다양한 정보를 일정한 이미지 위에 열 분포 형태의 비주얼한 그래픽으로 출력하는 것이 특징이다. X축과 Y축 변수 간의 연관성을 시각적으로, 수치로 확인할 수 있다. 값이 1에 가까이 갈 때는 X축과 Y축이 밀접한 연관이 있다고 볼 수 있다. 어느 정도의 수치가 되어야 밀접한 연관이 있다고 판단하는가의 기준은 주어져 있지 않지만, 히트맵에서 보이는 결과와 같이 온도, 강수량, 습도, 풍속의 변수 간의 상관관계는 밀접하게 관련이 되어 있

다고 볼 수 없다. 다만, 대여량과 관련하여 0.42를 보이는 온도와 0.48의 값을 보이는 시간에 대해서는 어느 정도 대여량과 상관관계가 있다고 볼 수 있다. 이 히트맵을 통해 대여량과 가장 연관성이 높은 변수는 시간(hour)과 온도(temperature)라는 것을 발견할 수 있다.

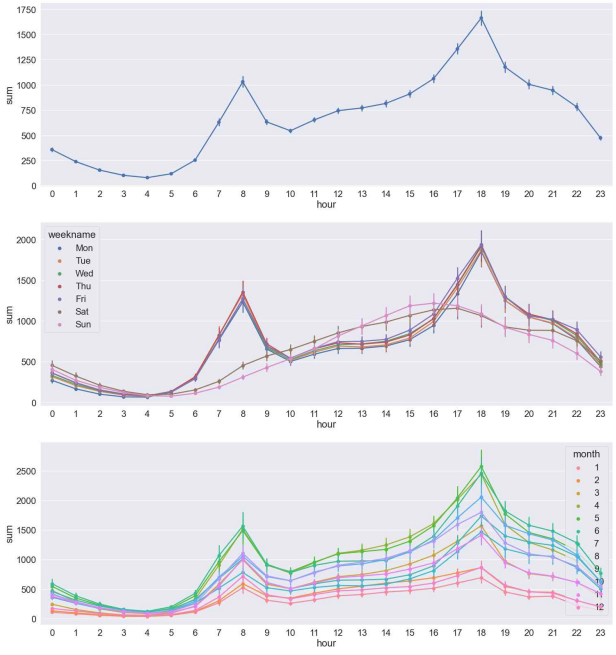


그림 12. 시간대별 대여량 추세
Fig. 12. Rental by hour

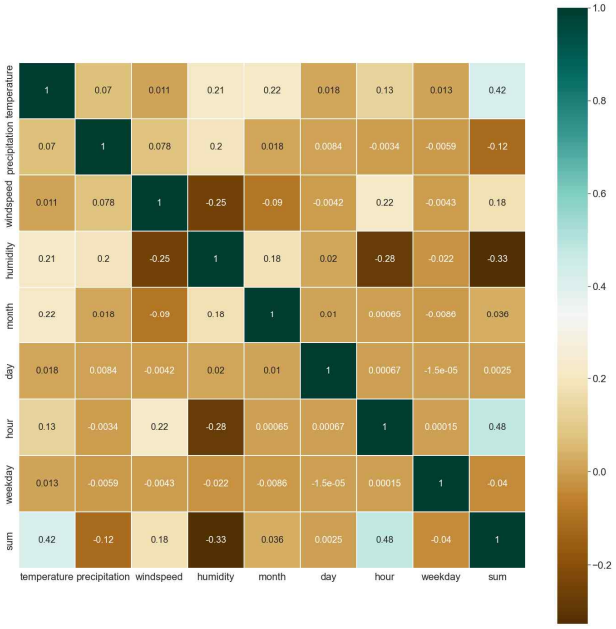


그림 13. 온도, 강수량, 습도, 풍속, 대여량의 상관관계
Fig. 13. Correlation by temperature, precipitation, humidity, windspeed, rental count

그림 14의 경우 연속형 변수의 관계를 파악하기 위해 온도, 강수량, 습도, 풍속, 대여량과의 산점도와 선형 회귀 적합선을 regplot을 통해 시각화하였다. 첫 번째 그림의 경우 온도가 영하와 같이 낮을 때는 대여량이 적고 온도가 높아질수록 대여량이 많아지지만 30도가 가까워지면 대여량이 점차 줄어드는 것을 확인할 수 있다.

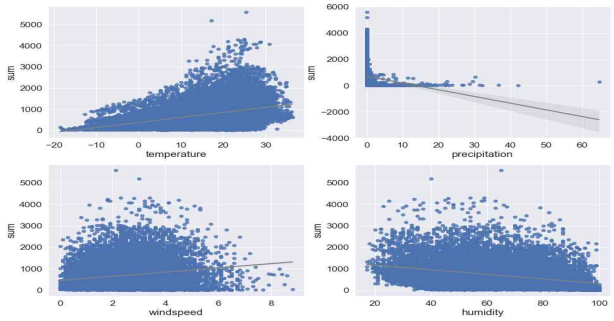


그림 14. 온도, 강수량, 풍속, 습도와 대여량의 산점도와 추세선
Fig. 14. Scatter plot and trend line by temperature, precipitation, windspeed, humidity

두 번째 강수량의 경우 결측치를 0으로 처리했기 때문에 상당수의 데이터가 0에 몰려있는 것을 볼 수 있다. 강수량의 데이터가 null인 이유는 관측되지 않은 값이라기보다 비가 오지 않았기 때문에 null 값이 있을 것으로 판단하여 0으로 처리하였다. 상대적으로 강수량 데이터가 있는 값이 많지 않아 그래프 자체로는 유의미한 정보로 표현되지는 않았으나 추세선과 산점도 그래프의 모양으로 보아 강수량이 많을 때 대여율이 적다는 것을 확인할 수 있다. precipitation(강수량) 그래프에서 확인할 수 있는 것 같이 실제 2021년 7월 19일 호우주의보가 있었던 날은 일 강수량이 67.4mm이었고 대여량은 0에 가깝다. precipitation 산점도 그래프의 Y축이 마이너스 -2000으로 표시된 것은 추세선 표시를 위해 seaborn 패키지에서 자동으로 표현된 것이고 추세선 표시를 하지 않으면 중심이 (0, 0) 인 그래프로 산점도가 그려진다.

세번째 windspeed(풍속) 산점도의 경우 풍속이 7 이상이면 대여량이 감소하는 것을 확인할 수 있으며, 네 번째 그림인 humidity(습도) 산점도의 경우 넓게 퍼져 있는 형태이지만 추세선의 기울기로 볼 때 습도가 높은 경우에는 대여량이 낮은 것을 확인할 수 있다.

Ⅳ. 모델링 및 평가

4-1 데이터 분할

데이터 분석에 사용한 최종 데이터 셋을 기준으로 학습 데이터와 테스트 데이터를 분리하였다. 대여량 예측 분석의 경우 과거의 데이터를 사용하여 훈련한 모델로 미래의 데이터를 예측하는 것이 효용성이 높으므로 2020년 7월 1일부터

2022년 6월 30일까지의 2년간의 최종 데이터 셋 17,435건 중 전년도 1년치인 2020년 7월 1일부터 2021년 6월 30일까지의 데이터를 학습 데이터로, 남은 2021년 7월 1일부터 2022년 6월 30일까지의 데이터를 테스트 데이터로 사용하였다. 통상적인 기계 학습 모델의 경우 전체 데이터의 80%를 모델 학습에, 20%를 테스트에 사용하는 것과 달리 데이터 셋을 분리한 이유는 자전거 대여 데이터 성격이 월별에 따라 달라지기 때문이다. 그래서 1년치 데이터를 학습 데이터로, 다른 1년치 데이터를 테스트 데이터로 사용한 것이다. 모델 학습을 위한 학습 데이터는 총 8,653건이며, 성능 평가에 사용하는 테스트 데이터는 8,782건이다.

예측 모델 성능을 높이기 위해 연속형 특징인 온도, 강수량, 습도, 풍속 정보와 범주형 특징인 년, 월, 일, 시간, 요일 정보의 타입을 카테고리 형으로 변경하였다.

4-2 모델링

기계 학습 라이브러리 중 사이킷런을 사용하고, 사이킷런에서 제공하는 지도학습의 분류 모델인 랜덤 포레스트, 그래디언트 부스팅, 결정 트리, 선형 회귀 모델을 사용하여 각각의 결과를 비교하였다. 학습 데이터로 해당 모델을 훈련하고 테스트 데이터로 모델을 평가하는 방식으로 진행하였다. 모델의 평가 지표는 MAE, MSE, RMSE, MAPE(mean absolute percentage error), RMSLE(root mean squared log error) 등 다양하게 존재하지만, 본 연구에서는 RMSE, R^2, RMSLE, 정확도 등 4가지의 지표로 각 모델을 비교 평가하겠다. RMSE는 예측 오차를 바탕으로 성능을 비교하며 오차의 측도로 사용하는 값이다. RMSE는 회귀 모형의 평가 지표로 많이 사용되며 RMSE가 낮은 모델이 예측을 더 잘한 것이다. 통계에서의 설명력을 보기 위해 R^2(결정계수, R Square)를 사용하는데 회귀 모델에서 예측 모델과 실제 모델이 얼마나 강한 상관관계가 있는가를 설명할 수 있다.

실제값과 예측값의 정확도 비율을 계산하기 위해 RMSLE 값을 사용하겠다. RMSLE를 사용하였을 경우 예측값이 1,500대이고 실제값은 1,000대였다면, 1.5인 값 즉 1.5배라고 부를 수 있는 값이 산출된다. 또한 직관적으로 성능을 평가하기 위해 정확도를 계산하였다. 예측한 결과와 실제값을 비교하여 정확도를 계산한 후 전체 합의 평균을 구한 값을 정확도로 사용하였다. 수식으로 나타내면 아래와 같다.

$$\text{정확도}(\%) = (1 - \Sigma(|\text{예측값} - \text{실제값}|/\text{실제값})/N) * 100$$

(단, |예측값-실제값|/실제값 > 1이면 최대 1로 설정)

실제값보다 예측값이 훨씬 큰 경우 오차의 비율이 100%를 넘어갈 수 있으므로 이럴 경우 최대 오차가 100%가 되도록 수식을 수정하였다. 예를 들어 실제값이 100이고 예측값이 150이라면 |예측값-실제값|/실제값 = |150-100|/100 = 0.5가 되지만, 실제값이 100이고 예측값이 300이라면 |예측값-실제값|/실제값 = |300-100|/100 = 2가 된다. 그러면 정확도의 계산값이 음수가 나올 수 있기 때문에 2를 1로 보정해서 계산했다. 즉, 실제값보다 예측값이 2배 이상 되면 오류는 100%, 정확도는 0%로 계산한 것이다.

1) 랜덤 포레스트(Random Forest)

랜덤 포레스트는 여러 개의 의사 결정 트리를 만들어 그들의 다수결로 결과를 결정하는 방식이다. 이 의사 결정 트리를 몇 개로 만드느냐가 n_estimators 값을 결정하지만 의사 결정 트리가 많다고 결과가 좋은 것은 아니고 트리가 많아지는 만큼 결과도 늦어지는 문제가 있다. n_estimators 최적의 값을 찾기 위해 10부터 2000까지의 값을 대비해 보았는데 n_estimators 값이 200일 때 가장 좋은 점수를 기록하다가 200 이후에는 점수가 낮아지는 이유로 n_estimators 값을 200으로 사용하였다.

예측한 결과는 RMSE 347.37, R^2는 0.74, RMSLE는 0.51이며, 정확도는 67.61%이다. 그림 15는 학습 데이터와 예측값을 비교하여 시각화한 결과이다. 두 그래프가 비슷한 형태를 나타내고 있는 것을 볼 수 있다.

2) 그래디언트 부스팅(Gradient Booting)

사이킷런의 GradientBoostingClassifier는 기본적으로 깊이가 3인 결정 트리를 100개 사용하며. 깊이가 얕은 결정 트리를 사용하기 때문에 과적합에 강하고 일반적으로 높은 성능을 기대할 수 있다.

부스팅 알고리즘은 여러 개의 약한 학습기(weak learner)를 순차적으로 학습, 예측하며 잘못 예측한 데이터에 가중치 부여를 통해 오류를 개선해 나가면서 학습하는 방식이다. 그래디언트 부스팅 가중치 업데이트를 경사 하강법(gradient descent)을 이용한다. 이 오류식을 최소화하는 방향성을 가지고 반복적으로 가중치의 값을 업데이트 해나가는 방식이다 [19].

그래디언트 부스팅에서도 n_estimators 인자가 사용되는데 이 값은 약한 학습기의 개수를 의미한다. 개수가 많을수록 예측 성능이 일정 수준까지는 좋아지지만, 계산량이 상대적으로 많아지기 때문에 수행 시간이 오래 걸리는 단점이 있다. 최적의 n_estimators 값을 찾기 위해 100 ~ 2000까지의 값을 대입해 보았고 1,000에서 가장 좋은 점수를 기록하여 n_estimators 값을 1,000으로 사용하였다.

예측한 결과는 다음과 같다. RMSE 356.18, R^2는 0.73, RMSLE 0.63으로 나왔으며 정확도는 63.60%이다. 그림 16은 학습 데이터와 비교하여 시각화한 그림이다. 예측 데이터 중 0 이상의 값에서 비슷한 형태로 보이나 자전거 대여량의 예측값이 0 이하의 값으로 표출된 것은 특이점으로 볼 수 있다.

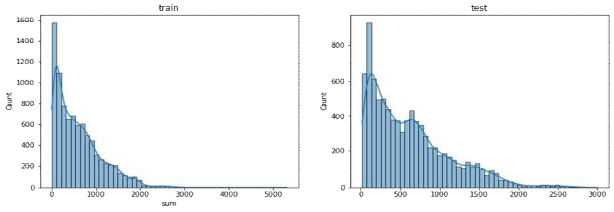


그림 15. 랜덤포레스트 모델의 학습 데이터와 예측값 비교
Fig. 15. Comparison of training data and predicted values of Random Forest model

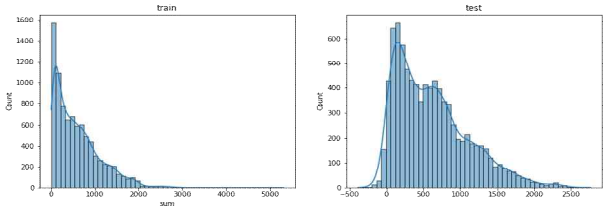


그림 16. 그래디언트 부스팅 모델의 학습 데이터와 예측값 비교
Fig. 16. Comparison of training data and predicted values of Gradient Boosting model

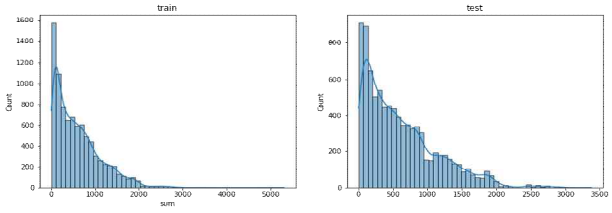


그림 17. 결정 트리 모델의 학습 데이터와 예측값 비교
Fig. 17. Comparison of training data and predicted values of Decision Tree model

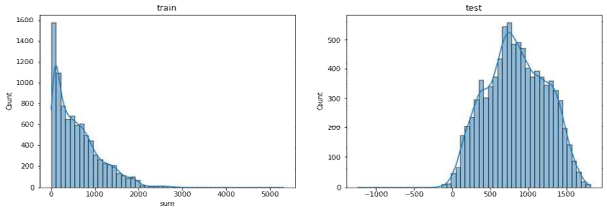


그림 18. 선형 회귀 모델의 학습 데이터와 예측값 비교
Fig. 18. Comparison of training data and predicted values of Linear Regression model

표 3. 평가표
Table 3. Evaluation table

Model / Result	Random Forest	Gradient Boosting	Decision Tree	Linear Regression
RMSE	347.37	356.18	438.66	500.08
R^2	0.74	0.73	0.59	0.47
RMSLE	0.51	0.63	0.74	0.93
Accuracy(%)	67.61	63.60	61.50	38.53

3) 결정 트리(Decision Tree)

결정 트리 생성 시 random_state 인자를 사용하게 되는데, 결정 트리를 분할해 나갈 때 특징 추출의 무작위 정도를 제어하기 위해 사용한다. 이 값이 None이면 실험할 때마다 특징이 임의로 추출된다. 본 연구에서는 일관성 있는 결과를 얻기 위해서 random_state 값을 0으로 세팅하였다.

예측한 결과는 RMSE 438.66, R^2는 0.59, RMSLE 0.74로 나왔으며 정확도는 61.60%이다. 그림 17과 같이 학습 데이터와 예측값을 비교하여 시각화하였을 때 전반적으로 비슷한 모양을 나타내고 있는 것을 볼 수 있다.

4) 선형 회귀(Linear Regression)

사이킷런은 선형 회귀 클래스로 선형 회귀 알고리즘을 구현했기 때문에 해당 클래스의 객체를 만들어 훈련, 평가, 예측할 수 있으며 다른 모델 역시 사이킷런의 모델 클래스의 fit(), score(), predict() 동일한 메서드 이름을 사용한다.

예측한 결과는 RMSE 500.27, R^2는 0.47, RMSLE 0.93로 나왔으며 정확도는 38.53%이다. 그림 18은 학습 데이터와 예측값을 비교하여 시각화하였을 때 예측값이 0 이하의 값이 표출되었고 실제값과 예측값의 그래프를 보았을 때 모양의 형태가 전혀 다른 것을 알 수 있다.

4-3 평가

기계 학습 알고리즘 분석 결과 시간 정보와 기후 정보 전체를 변수로 사용하였으며 서울 전 지역을 대상으로 분석하여 RMSE가 상대적으로 크게 나온 것으로 보인다. 이를 보완하기 위해 통계의 결정계수 R^2를 산출하여 성능을 비교하였다. 또한, 실제값과 예측값의 정확도 비율을 구하기 위해 RMSLE 값을 구해 보았다. 예측과 실제값의 평균 비율을 구할 수 있다. RMSLE 가 0이라면 실제값과 예측치가 같은 경우이므로 숫자가 0에 가까울수록 예측 정확도가 높다고 볼 수 있다. 정확도는 100%가 가장 정확한 결과이며 실제값과 예상치가 2배 이상의 오차가 발생할 때는 정확도를 0%로 계산하였다.

예측 모델에서 예측한 값과 실제 값 사이의 평균 차이를 측정하는 RMSE의 경우에는 랜덤 포레스트와 그래디언트 부스팅 간에는 큰 차이를 보이지 않지만 결정 트리, 선형 회귀와는 상당한 차이를 보여준다. 결정계수 R^2의 경우에도 랜덤 포레스트와 그래디언트 부스팅 간에는 큰 차이를 보이지 않지만 결정 트리, 선형 회귀는 점수에서 상당한 격차를 보인다.

MSLE 역시 예측값과 실제 값의 차이를 비교한 것으로 랜덤 포레스트가 가장 높은 결과를 나타냈고 선형 회귀는 가장 낮은 결과를 나타내었다. 정확도의 경우 랜덤 포레스트가 가장 높은 결과를 나타내었다. 랜덤 포레스트, 그래디언트 부스팅, 결정 트리 모두 60점대를 기록했지만 선형 회귀의 경우 38.53으로 낮은 수치를 기록하였다.

선형 회귀는 종속 변수와 한 개 이상의 독립변수 간의 선형 상관 관계를 모델링하는 회귀분석 기법으로 기계 학습 알고

리즘 가운데 가장 많이 사용되는 기법이긴 하나, 종속 변수와 독립 변수 간에 선형 방식이 아닐 경우 최적의 접근 방식이 아닐 수 있다. 대여량인 종속 변수와 시간, 요일, 날씨와 같은 독립변수 간의 관계가 선형이 아닐 수 있기 때문에 대여량 예측에 낮은 평가를 기록한 것으로 판단된다.

RMSE, R^2, RMSLE, 정확도(%)의 전체적인 결과를 보면 랜덤 포레스트, 그래디언트 부스팅, 결정 트리, 선형 회귀 순으로 평가 결과를 나타내었다. 전반적으로 랜덤 포레스트, 그래디언트 부스팅, 결정 트리 모델이 선형 회귀에 비해 좋은 성능을 보여주며 표 3에서 보는 것과 같이 최종적으로는 랜덤 포레스트 모델이 다른 알고리즘보다 RMSE, R^2, RMSLE, 정확도(%)의 종합적인 결과에서 가장 성능이 좋은 것으로 나타났다.

V. 결 론

본 연구에서는 사이킷런에서 제공하는 기계 학습의 다양한 기법을 사용하여 서울시 공공자전거 대여량을 예측하고 정확도를 측정하여 어떤 것이 가장 좋은 성능을 보이는지 확인하고 비교 분석하는 것에 의의가 있다. 결과를 도출하기 위하여 공공자전거 대여량 데이터 자료뿐만 아니라 기상청 날씨 자료를 포함하여 데이터셋을 구성하였다. 데이터의 이해를 위해 분석 작업 및 시각화를 진행하였고 데이터의 패턴과 대여량과의 관계들을 시각적으로 확인할 수 있었다.

분석을 위해 수집한 2년 치 데이터 중 20%를 비복원 추출하여 17,435건의 최종 데이터셋을 만들어 학습 데이터와 테스트 데이터로 구성하였다. 과거의 데이터를 분석하여 미래의 데이터를 예측한다는 측면으로, 2년 치 데이터 중 전년도 데이터는 학습 데이터로 사용하였고, 다음 해 데이터는 테스트 데이터로 사용하였다.

사이킷런에서 제공하는 지도학습 모델인 랜덤 포레스트, 그래디언트 부스팅, 결정 트리, 선형 회귀 등 4개의 모델을 사용하여 분석을 수행했다. 학습 데이터로 해당 모델을 훈련하고 테스트 데이터로 모델을 평가하는 방식으로 진행하였다. 각 모델별 최상의 성능으로 결과를 도출하기 위해 의사 결정 트리의 최적의 인자값을 추출하여 학습 과정에 사용하였다.

예측의 오차를 비교 평가하기 위해 RMSE, R^2, RMSLE, 정확도(%)를 계산하여 평가 방법으로 사용하였다. 그 결과 RMSE, R^2, RMSLE, 정확도(%) 전체적인 측면에서 랜덤 포레스트, 그래디언트 부스팅, 결정 트리, 선형 회귀 순으로 예측 성능이 높게 평가되었다. 최종적으로 랜덤 포레스트 모델이 가장 좋은 성능을 보여주었다. 이러한 결과를 활용하여 지역별 이용 현황 및 대여소별 이용 건수 예측, 특정 시간대별 이용 건수 예측 등 분석 목적에 따른 다양한 조건 부여로 세부적인 대여량 예측을 통한 공공 자전거 운영에 효율성을 줄 뿐만 아니라 이를 활용한 마케팅 및 비즈니스 전반에 도움을 줄 수 있을 것으로 보인다.

본 연구의 예측 분석에 사용한 특징(feature)은 온도, 강수량, 습도, 풍속, 년, 월, 일, 시간, 요일 정보이지만 실제 대여량에 영향을 미칠만한 변수들은 다양하게 존재하고 있다. 향후 대여량과 관련성 있는 대여소의 위치, 이동 거리, 미세 먼지, 축제와 같은 이벤트 여부, 코로나19와 같은 사회적 이슈, 공휴일 등의 변수들을 추가하거나 기계 학습의 다양한 모델과 파라미터(parameter)의 최적화를 통해 예측 분석을 적용해 본다면 더욱 개선된 예측 결과를 도출할 수 있을 것으로 생각된다.

참고문헌

[1] Seoul Metropolitan Government Bicycle Policy Division. Bicycle Policy Team Notice (January 2021) [Internet]. Available: https://news.seoul.go.kr/traffic/archives/504919?listPage=2&s=%EB%94%B0%EB%A6%89%EC%9D%B4

[2] Seoul Metropolitan Government Bicycle Policy Division. Public Bike Team News Report (June 2022) [Internet]. Available: https://www.seoul.go.kr/news/news_report.do#view/366295

[3] H. S. Park, *Machine Learning + Deep Learning To Study Alone*, HANBIT Media, 2020.

[4] G. H. Lee and S. K. Song, "Predicting Flight Delays Based on Deep Neural Network," in *Proceedings of Korea Computer Congress 2021*, Jeju, pp. 1075-1076, June 2021.

[5] M. Y. Kang and S. Kim, "A Study on Production Prediction Model Using a Energy Big Data Based on Machine Learning," in *Proceedings of the 2022 Korean Institute of Information and Commucation Sciences Autumn Conference*, Jeju, Vol. 26, No. 3, pp. 453-456, October 2022.

[6] J. Y. Choi, H. Y. Yang, and H. Y. Oh, "Store Sales Prediction Using Gradient Boosting Model," *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 25, No. 2, pp. 171-177, February 2021. https://doi.org/10.6109/jkiice.2021.25.2.171

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, ... and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, Vol. 12, pp. 2825-2830, November 2011. https://doi.org/10.48550/arXiv.1201.0490

[8] A. Fandango, *Python Data Analysis − Second Edition*, Seoul: Acorn Publishing, 2018.

[9] S. Song and H. Lee, *Data Analysis for Everyone: With Python*, Seoul: Gilbut, 2019.

[10] Seoul Metropolitan Government. Seoul Open Data Plaza [Internet]. Available: https://data.seoul.go.kr/.

[11] Korea Meteorological Administration. Open MET Data

Portal [Internet]. Available: https://data.kma.go.kr/.
[12] DBpia. Academic Information Portal [Internet]. Available: https://www.dbpia.co.kr/.
[13] J. W. Min, H. S. Mun, and Y. S. Lee, “Demand Forecast for Public Bicycles (‘Tashu’) in Daejeon Using Random Forest,” in *Proceedings of the Korean Information Science Society Conference*, Jeju, pp. 969-971, June 2017.
[14] K.-M. Cho, S.-S. Lee, and D. Nam, “Forecasting of Rental Demand for Public Bicycles using a Deep Learning Model,” *The Journal of the Korea Institute of Intelligent Transport Systems*, Vol. 19, No. 3, pp. 28-37, June 2020. https://doi.org/10.12815/kits.2020.19.3.28
[15] S. Min and Y. Jung, “Comparative Study of Prediction Models for Public Bicycle Demand in Seoul,” *Journal of the Korean Data and Information Science Society*, Vol. 32, No. 3, pp. 585-592, May 2021. https://doi.org/10.7465/jkdi.2021.32.3.585
[16] Towards Data Science. Data Science Trends Based on 4 Years of Kaggle Surveys [Internet]. Available: https://towardsdatascience.com/data-science-trends-based-on-4-years-of-kaggle-surveys-60878d68551f.
[17] Y. Kang, D. Park, and S. Kim, *Best Machine Learning*, Paju: Life & Power Press, 2021.
[18] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python*, 2nd ed. Seoul: Hanbit Media, 2022.
[19] C. Kwon, *The Ultimate Guide to Python Machine Learning*, Paju: Wikibooks, 2019.

권혜진(Hye-Jin Kwon)

2023년 : 강원대학교 컴퓨터정보통신공학과 (공학석사)
2008년～2019년: 서울시청
※관심분야 : 인공지능, 기계 학습, 데이터분석

하진영(Jin-Young Ha)

1987년 : 서울대학교 컴퓨터공학과 (공학사)
1989년 : KAIST 전산학과 (공학석사-인공지능)
1994년 : KAIST 전산학과 (공학박사-인공지능)
1994년～1997년: ㈜ 핸디소프트 기술연구소 책임연구원
2000년～2001년: IBM T. J. Watson Research Center 방문연구원
1997년～현 재: 강원대학교 컴퓨터공학과 교수
※관심분야 : 인공지능, 패턴인식, HCI