결과 보고서

소속	경북대학교		
팀 명	2419		
팀 원	권병근, 박민서, 성다솜, 이다인		
제 목	XGBoost, LightGBM, CatBoost 앙상블 모델과 SHAP_Value를 활용한 전력기상지수 예측 및 주요 변수 분석		
과제명	공모분야: 기상에 따른 공동주택 전력수요 예측 개선		
활용 도구	작업환경	Jupyter Notebook (Anaconda3), VSCode Python version: 3.11.3	
	라이브러리	pandas (2.2.1) matplotlib (3.8.3) scikit-learn (1.4.1) tqdm(4.66.4) numpy(1.26.4) holidays(0.49)	geopy(2.4.1) pyproj(3.6.1) xgboost (2.0.3) lightgbm (4.3.0) catboost (1.2.5) shap (0.45.1)

Introduction

본 과제의 목표는 기상 변수 및 공공 데이터를 활용하여 공동주택 전력 수요 증강에 영향을 미치는 요인을 분석하고, 계절 및 지역에 따른 모델 세분화를 통해 공동주택 전력 수요를 예측(전력기상지수)하는 최적의 모델을 개발하는 것이다. 전력기상지수는 기상 변화에 따른 지역별 공동주택의 예상 전력 부하 변화를 기상 예보처럼 국민들이 쉽게 이해할 수 있도록 수치화 하여 예측하는 서비스이다.

본 팀은 전력기상지수를 예측하기 위해 다양한 파생 변수를 생성하고, 여러 머신러닝 모델 앙상블을 통해 일반화 성능을 향상시켰다. 모델 구축 후 각 변수가 결과값에 미치는 영향을 파악하기 위해 SHAP Value 분석 기법을 사용했다. 모델 평가 결과, 전력 기상지수 예측값과 실제 전력 수요 간의 상관관계가 0.977로 높은 성능을 보였다. 전력기상지수를 결정 짓는데 가장 큰 영향을 준 것은 시간대, 체감 온도, 최저기온, 휴일 여부, 불쾌지수 등이 있다. 해당 변수들의 분석 결과, 사람들이 주택에 오래 머무를 조건에서 전력기상지수가 상승할 확률이 높아지는 것을 확인했다. 또한, 불쾌지수가 상승하면 전력기상지수도 높아지는 경향을 보인다. 이는 여름철에 냉방장치 사용량이 전력기상지수 상승에 큰 영향을 미친다는 것을 의미한다.

Data Preprocessing

- 데이터 정보

electric_train.csv 파일에는 총 7,593,355개의 행과 16개의 열로 데이터가 구성되어 있다. 각 격자 번호에 대해 2020년부터 2022년까지의 연도 중 최대 3년치의 데이터가 1시간 단위로 행이 구성되어

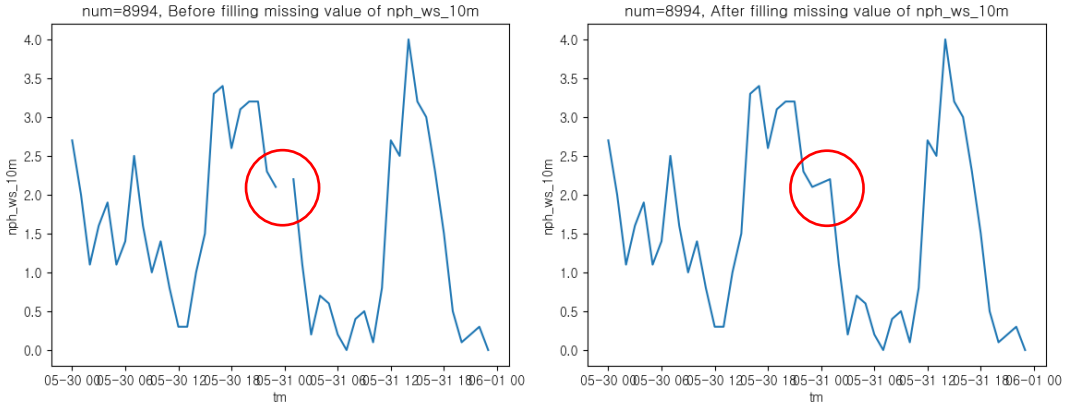
있다. 또한, train과 test 파일의 모든 열의 이름에 'electric_train', 'electric_test'와 같이 파일 명이 포함되어 있어 편의성을 위해 제거하여 열 이름을 간단히 변환했다.

열 이름	내용	열 이름	내용
num	동네예보 격자번호	nph_ta	기온
tm	전력부하 측정 시간 (시간 포함)	nph_hm	상대습도
hh24	전력부하 측정 시간 (1~24)	nph_ws_10m	10분 평균 풍속
n	공동주택 수	nph_rn_60m	1시간 누적 강수량
stn	AWS 지점 번호	nph_ta_chi	체감온도
sum_qctr	계약전력합계	weekday	요일
sum_load	전력수요 합계	week_name	주중 주말
n_mean_load	전력부하량 평균	elec	전력기상지수 (타겟 변수)

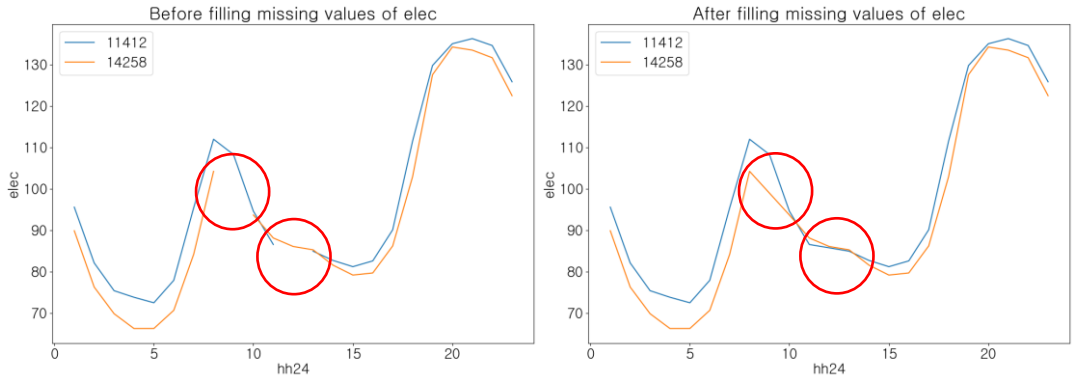
[표 1] elec_train.csv 열 구성 목록

- 결측치 처리

train파일의 'nph_ws_10m'열과 train 파일의 타겟 변수인 'elec'열에서 -99로 채워진 결측치가 있었다. 'nph_ws_10m' 열에서는 총 169개의 행에서 결측치가, 'elec'열에서는 총 5개의 행에서 결측치가 있었다. 이들에 대한 결측치들은 linear interpolation으로 채웠다.



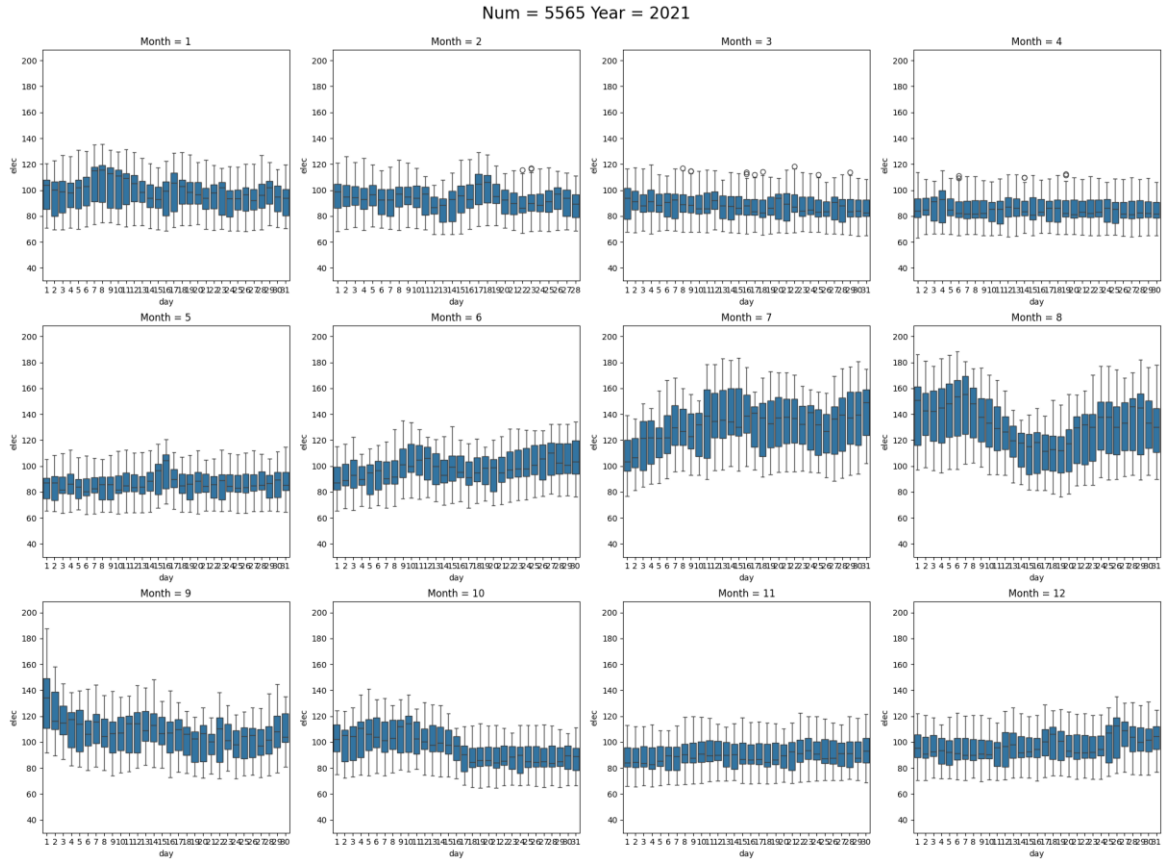
[그림 1] nph_ws_10m 결측치 채우기 전과 후



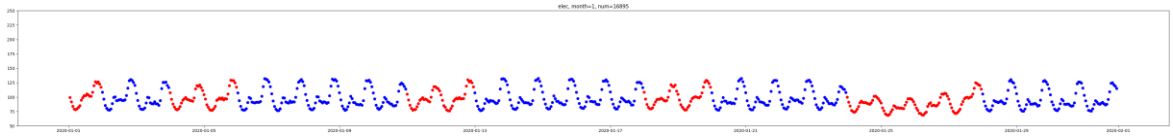
[그림 2] elec 결측치 채우기 전과 후

- 탐색적 데이터 분석(EDA)

타겟 변수인 'elec'에 대해 월별 Boxplot을 그려본 결과 몇 가지 사실들을 알아낼 수 있었다. 우선, 대부분의 기간에서 비슷한 수준의 값을 유지하다가 6월에 들어서면서 일정 수준을 뛰어넘고 9월과 10월 즈음에 값이 일정 수준으로 되돌아오는 것을 관찰할 수 있었다. 또한, 평일과 주말 혹은 공휴일 간의 elec 값의 평균과 시간 별 패턴이 차이가 있음을 확인할 수 있었다. 이를 통해, 사람들이 주택에서 여름에 대비하기 위해 에어컨과 선풍기와 같은 냉방 장치들의 사용량이 증가하여 전력지수가 평소보다 해당 기간에 더 큰 값을 가진다는 것과 주택에 있는 시간이 평일보다 주말과 공휴일에 더 많을것이란 가정을 할 수 있다. 또한, 지역권별로 'elec' 값의 평균, 격자 내 공동주택 수, 전력 수요량이 다르다는 것을 확인할 수 있었다. 이 점을 미루어 보아 전력 지수에 지역의 특성이 어느 정도 영향이 있다고 볼 수 있다. 해당 가정들을 통해 시간과 기상, 위치와 관련된 파생변수들을 생성했다.



[그림 3] 한 격자 내의 월별 elec값 Box plot



[그림 4] 16895번 격자 내의 1월 elec값의 변화

- 파생변수

1. 시간 데이터 (month, day_of_year, rest_day)

기존에 주어진 데이터 속 시간에 관련된 데이터를 잘 처리할 수 있게 'tm'열 속에 있는 년, 월, 일

'year', 'month', 'day'열로 분리하여 저장했다. 또한, 매년 1월 1일부터 12월 31일까지 1부터 365(혹은 366, 윤년)까지의 번호를 부여하여 'day_of_year'라는 파생변수를 생성하며 날짜 계산을 효율적으로 가능하게 했다. 기존의 'week_name' 변수에 대해서 주중과 주말에 대해서만 반영되어 있기 때문에, 주말과 공휴일을 같이 엮어 쉬는 날과 그렇지 않은 날을 1과 0으로 구분하여 저장한 'rest_day'를 생성하여 공휴일을 반영하고자 했다.

2. 기상 관련 변수(ta_diff, discomfort, max_temp, min_temp, CDH_26, HDH_18)
기상 현상 중 기온에 관련하여 변수들을 생성하고자 하였다. 우선, 실제 기온에서 체감 온도를 뺀 값을 'ta_diff'라는 변수와 주어진 기온과 상대습도를 이용한 불쾌지수(discomfort)를 계산하여 사람이 직접 느끼는 온도를 반영하고자 하였다. 불쾌지수 계산 식은 다음과 같다.

$$Discomfort = \frac{9}{5}(ta) - 0.55(1 - hm)\left(\frac{9}{5}ta - 26\right) + 32$$

또한, 각 격자 별로 하루 최고 기온과 최저 기온을 찾아 'max_temp'와 'min_temp'에 저장했다. 그리고, 냉방 장치 및 난방 장치와 연관이 있는 냉방도시간(cooling degree hours)와 난방도시(heating degree hours)를 계산하여 'CDH_26', 'HDH_18'에 저장하였다. 이 변수들은 하루 동안 기온이 냉방 기준온도인 26도보다 높은 시간, 혹은 난방 기준 온도인 18도보다 낮은 시간을 계산하여 저장했다.

3. 위치 관련 변수(grid_x, grid_y, location_num)
'기상자료개방포털'에 공개되어 있는 자료인 '동네예보 영역 및 격자점 정보'를 참고했다. 격자들은 가로 149개, 세로 253개로 이루어져 있고, 주어진 자료들의 num열을 통해 격자점의 좌표들을 계산하여 'grid_x'와 'grid_y'열에 저장했다. 그리고, 해당 자료 속 소개되어 있는 지도투영법(Lambert-Conformal map projection)을 geopy와 pyproj 모듈을 이용하여 해당 격자의 경도와 위도를 얻어낼 수 있었다. 이를 통해, 격자가 속한 지역을 반환할 수 있었고, 해당 지역을 광역시와 전국 8도로 분류하여 번호를 지정하여 'location_num'열에 저장했다. 번호 지정 시에는 이전 주민등록번호 생성 시 활용되었던 지역번호를 이용했다.

Method

데이터는 시계열 형태를 띠고 있었으나, 과거 변수들이 현재에 미치는 영향이 크지 않다고 판단하여 다중 회귀 분석 모델을 활용했다. 머신러닝에서는 주로 사용되는 XGBoost(eXtreme Gradient Boosting), LightGBM(Light Gradient Boosting Machine), CatBoost(Categorical Boosting), RandomForest 등의 모델과 OLS, Ridge, Lasso 등의 전통적인 통계 회귀 기법을 활용하여 실험을 진행했다. 그 결과, 성능이 가장 우수했던 XGBoost, LightGBM, CatBoost 세 가지 모델을 최종적으로 채택했다. XGBoost는 강력하고 효율적인 구현으로 널리 사용되는 Gradient Boosting 프레임워크이다. 뛰어난 성능과 확장성을 자랑하며, 정교한 트리 기반 학습 알고리즘을 통해 높은 예측 정확도를 제공한다. 주로 과적합을 방지하기 위한 규제(regularization) 기법을 도입하여 다양한 데이터셋에서 우수한 성능을 발휘한다. LightGBM은 Microsoft에서 개발한 Gradient Boosting 프레임워크로, 대규모 데이터셋과 높은 속도의 학습 및 예측을 위해 설계됐다. Leaf-wise 트리 분할 방식을 사용하여 학습 시간을 단축하고 메모리 사용량을 줄이는 동시에, 높은 예측 성능을 유지한다. 특히 대용량 데이터와 고차원 특징을 다루는 데 효율적이다. CatBoost는 Yandex에서 개발한 Gradient Boosting 알고

리즘으로, 범주형 변수를 효과적으로 처리하는 데 강점을 가지고 있다. 데이터의 순서에 민감하지 않으며, 고유의 범주형 데이터 인코딩 기법을 통해 예측 성능을 높인다. 또한 과적합을 방지하기 위한 다양한 기법을 포함하여, 다양한 실전 문제에서 안정적이고 높은 성능을 발휘한다.
각 모델은 개별적으로도 우수한 성능을 보였으나, Soft Voting 을 활용한 앙상블을 통해 기존보다 더 일반화 된 모델을 개발했다. 앙상블 기법의 장점은 다음과 같다. 첫째, 여러 모델의 예측을 결합함으로써 단일 모델의 약점을 보완할 수 있다. 둘째, 다양한 모델이 상호 보완적인 역할을 하여 예측의 안정성과 정확성을 높인다. 셋째, 개별 모델의 과적합을 줄여 일반화 성능을 개선한다. 이러한 앙상블 기법의 장점을 통해 전력기상지수 예측의 신뢰성과 정확도를 극대화했다.
실험에 앞서 기존의 Train 데이터 셋을 X_train 데이터 셋과 X_test 데이터 셋으로 분리하는 train_test split의 비율을 0.2로, 모델의 모든 random_state 값은 42로 고정했다.

Conclusion

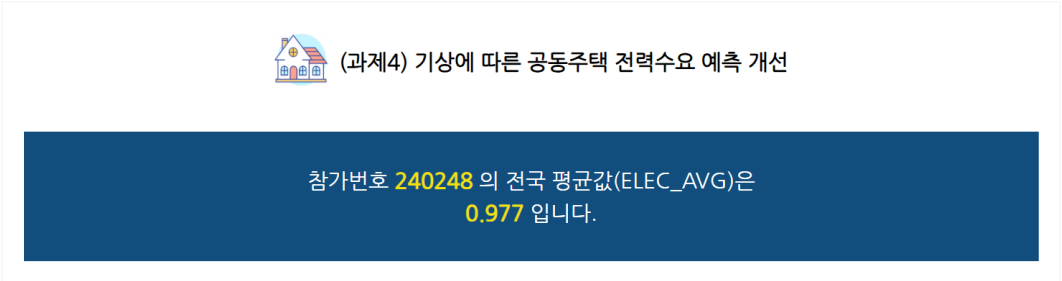
- 실험

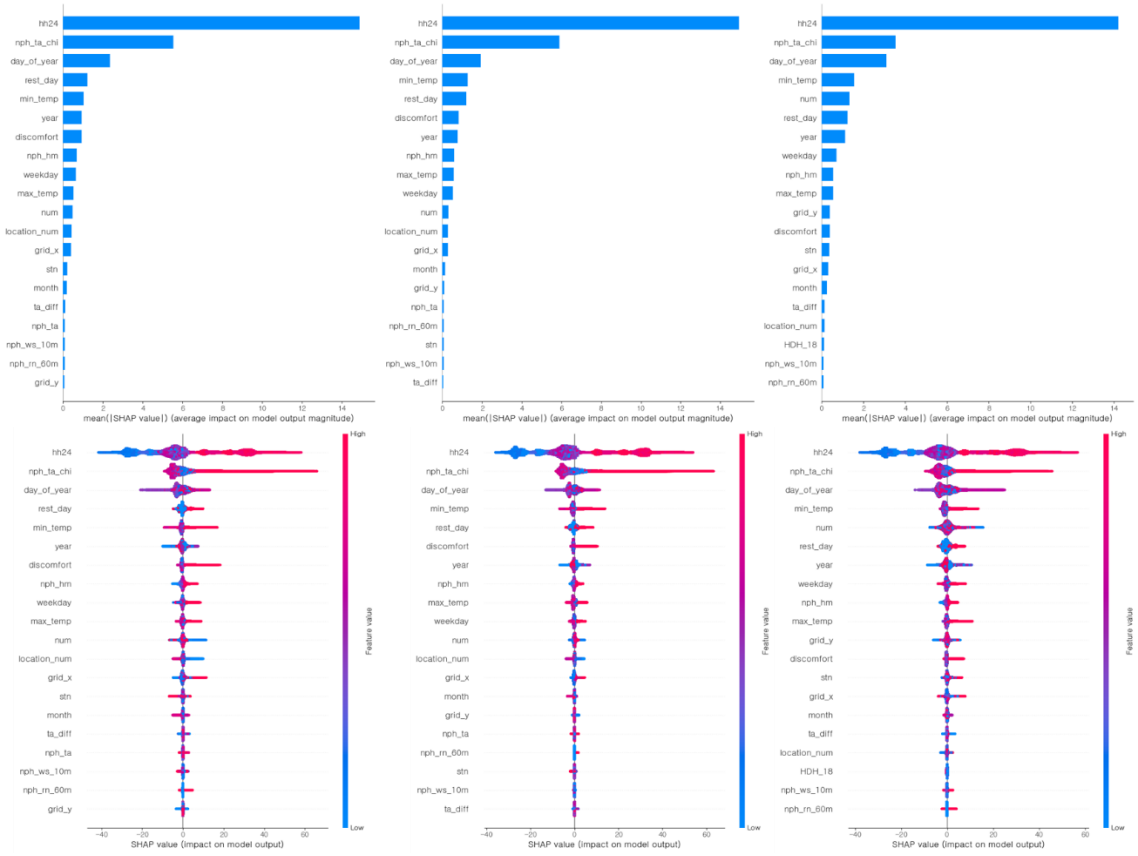
채택한 XGBoost, LGBM, Catboost 모델로 단일 모델 및 각각을 앙상블한 모델을 생성해 동일한 조건 하에 실험을 진행했다. 실험의 결과는 다음과 같다.

	RMSE		Total Correlation	
Model	X_train	X_test	X_train	X_test
XGB	4.468	4.490	0.9840	0.9839
LGBM	5.241	5.252	0.9780	0.9779
Catboost	3.680	3.696	0.9892	0.9891
XGB + LGBM	4.676	4.691	0.9825	0.9824
XGB + Cat	3.888	3.906	0.9879	0.9878
LGBM + Cat	4.221	4.233	0.9858	0.9857
XGB + LGBM + Cat	4.192	4.207	0.9860	0.9859

[표 2] 실험 결과

단일 모델링 결과 보다는 앙상블 모델들에서 높은 성능을 보여주었다. 자체 평가에서는 XGB + Cat 모델이 가장 좋았지만, 실제 홈페이지 검증 결과에서는 XGB + LGBM + Cat 앙상블 모델이 더 높은 수치를 보였다. 따라서 더 일반화된 모델을 선정하기 위해 XGB + LGBM + Cat 앙상블 모델을 채택했다.





XGBoost LightGBM CatBoost

앙상블 모델에서 각각의 변수들이 모델에 어떻게 영향을 미치는지 파악하기 위해 SHAP Value 분석을 진행했다. 분석 결과, 전력기상지수에 큰 영향을 미치는 주요 변수로는 hh24(24시 시간대), nph_ta_chi(체감온도), min_temp(최저기온), rest_day(휴일 여부), discomfort(불쾌지수)가 확인되었다. 먼저, hh24와 rest_day를 통해 사람들의 주택 거주 시간이 길어지는 저녁 시간대와 휴일일수록 전력기상지수가 높아진다는 것을 알 수 있다. 이는 사람들이 주로 집에 머무는 시간대와 휴일에 에너지 사용량이 증가하는 경향을 반영한다. 또한, nph_ta_chi와 min_temp의 경우, 온도가 높아질수록 전력기상지수가 증가하는 현상을 확인할 수 있다. 이는 여름철에 냉방장치 사용량이 많아짐에 따라 전력 사용량이 증가하는 것을 의미한다. 특히 체감온도와 최저기온이 높을수록 사람들은 더 많은 냉방을 필요로 하게 되어, 전력 소비가 크게 늘어나는 양상을 보인다. 마지막으로 discomfort는 기온과 상대습도에 의존하는 변수이다. 기온과 상대습도가 상승하면 discomfort 역시 상승하게 되며, 이는 전력기상지수 상승에 영향을 미친다. 특히 장마철에는 불쾌지수가 높아져 냉방장치 사용량이 급증하게 되고, 이는 전력 사용량의 증가로 이어진다. 결론적으로, 이 분석을 통해 저녁 시간대와 휴일, 그리고 여름철 기온 상승과 높은 불쾌지수가 전력 사용량 증가에 주요한 영향을 미친다는 것을 확인할 수 있었다. 이러한 결과는 에너지 수요 관리와 효율적인 전력 공급 계획 수립에 중요한 기초 자료로 활용될 수 있을 것이다.