

DACON AI 경진대회 Mini – Project

2025 전력사용량 예측

프로젝트 요약

안정적이고 효율적인 에너지 공급을 위해 전력 사용량 예측의 중요성은 갈수록 커지고 있습니다. 기후 변화와 에너지 전환 정책이 가속화되며, 수요 예측 기반의 에너지 관리 역량은 산업 전반의 핵심 경쟁력으로 자리잡고 있습니다. 따라서 본 프로젝트에서는 기상정보 및 건물 특성을 바탕으로 시간 단위 전력 수요를 예측함으로써, 건물별 에너지 관리 최적화 및 수요반응 기반 운영 정책수립에 기여하는 것을 목표로 합니다. 타겟변수인 전력소비량과 유의한 중요 피처들을 가설을 세워 통계적으로 분석하고 결측치, 이상치, 모델 성능 향상을 위한 데이터 정규화 및 변환의 데이터 전처리 과정을 이후 머신러닝 모델링 비교를 통해 전력 사용량을 예측하였습니다.

PROCESS

1. 문제 정의

- ✓ 배경
- ✓ 가설 설정

2. 탐색적 데이터 분석 및 전처리

- ✓ EDA
- ✓ 데이터 전처리
- ✓ 최종 변수 선택

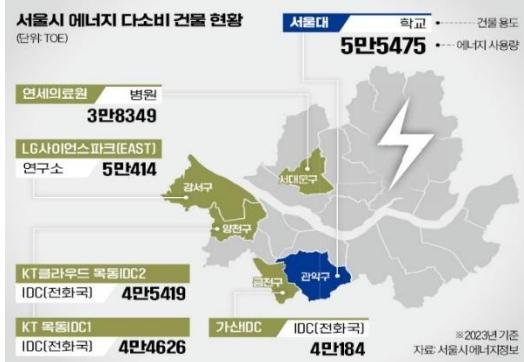
3. 데이터 모델링

- ✓ 모델 학습 비교
- ✓ 모델 예측 및 평가
- ✓ 모델 변수 중요도

4. 결론

- ✓ 기대효과
- ✓ 팀원별 회고

배경



건물 특성에 따른 전력소비 영향

기온에 따른 전력소비 영향

신재생에너지
전력소비량 영향

건물유형, 시간대, 기상 조건 등 복합적인 변수에 따라 전력 수요는 시시각각 변화하며
안정적인 에너지 공급 및 효율적인 수요 관리를 위해 머신 러닝을 활용한 예측이 필요하다

통계적인 가설 수립

1

귀무가설(H_0)

건물 유형에 따라 전력소비량의 평균은 차이가 없다.

2

대립가설(H_1)

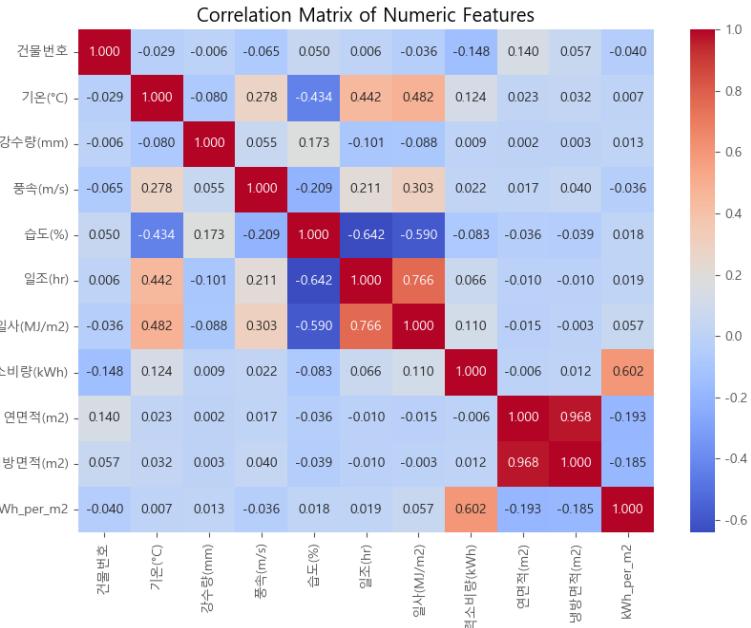
건물 유형에 따라 전력소비량의 평균은 차이가 있다.



추가 가설

기온, 냉방면적, 태양광용량은 전력 소비량 분류에 유의미한 영향을 미친다.

상관관계 분석



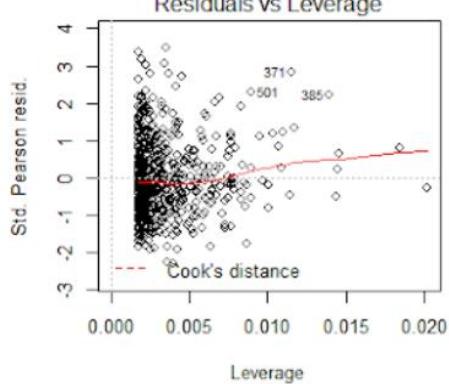
전력소비량은 건물유형, 시간, 기후 요인이
복합적으로 영향을 미치는 것으로 분석

변수선택 고려

데이터의 정규성 및 등분산성 확인



Std. Pearson resid.



데이터 비정규성 및 이분산성으로 가설에
대한 비모수 검정 실시, 각각의 변수에 대한
상관성 고려

시간 관련 분석

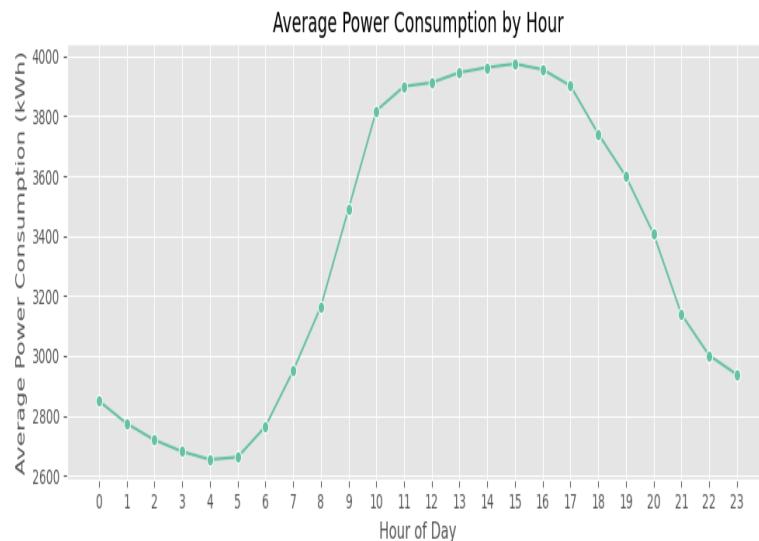
00:00 ~ 05:00 대부분 활동이 중단된 시간으로 소비량이 낮음 : 평균 2600Kwh

10:00 ~ 16:00 활동 시간에 따라 전력 또한 최고 피크타임 기록 : 평균 4000Kwh

17:00 ~ 23:00 업무 종료에 따른 전력감소 구간 : 최저 2900kwh



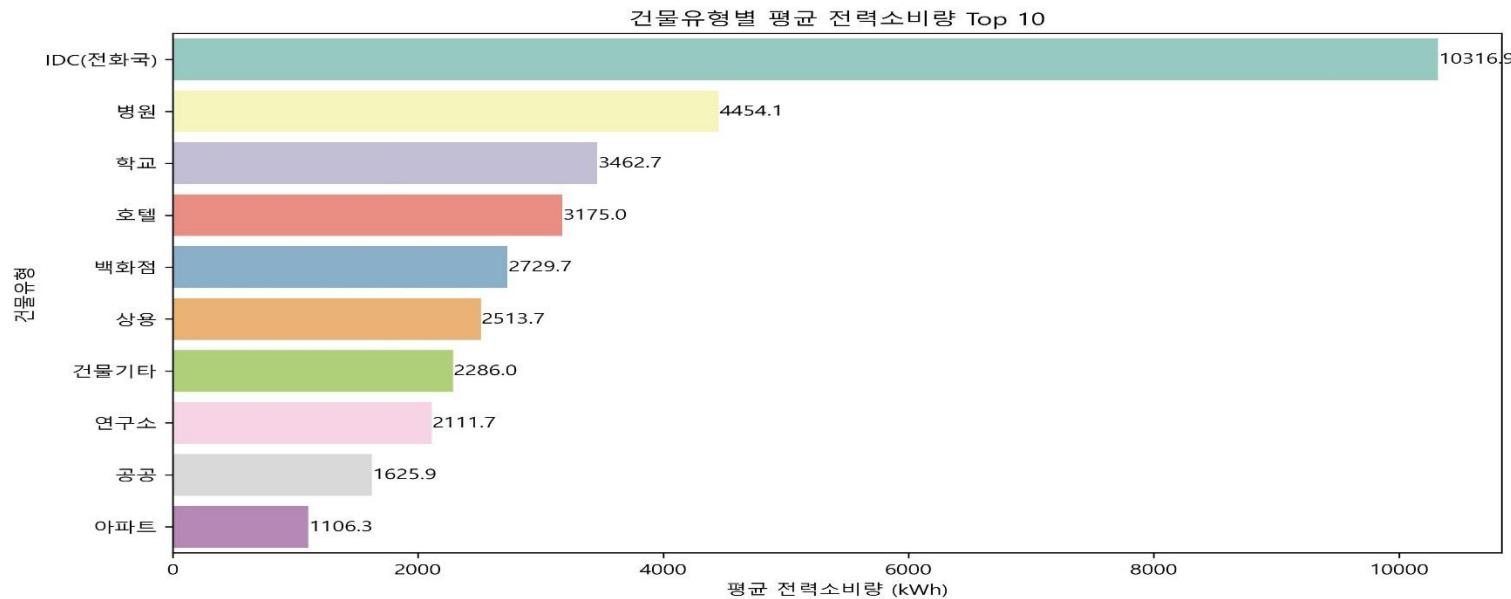
사람들의 활동 시간에 따라 전력 또한 상승 및 하강구간이 존재한다.



건물 유형별 평균 전력소비량

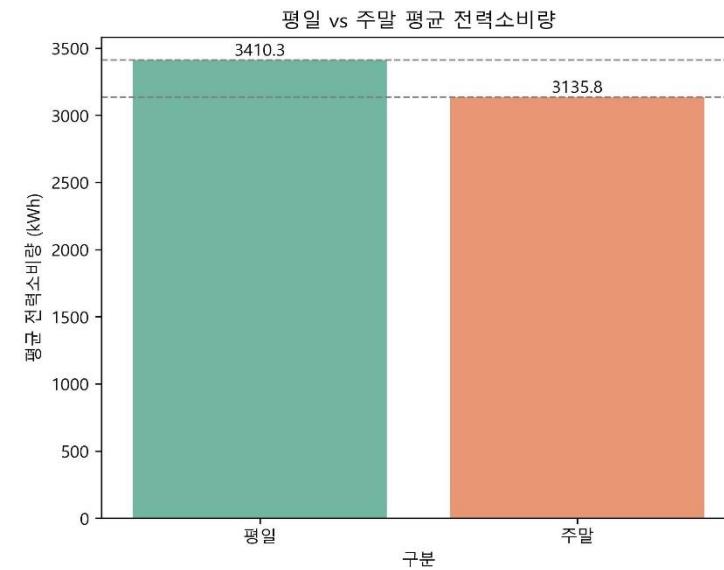
전화국은 병원보다 평균전력을 130% 더 소비하고

병원, 학교, 호텔 순으로 건물 유형별 평균 약 10% 전력소비량 차이가 있다

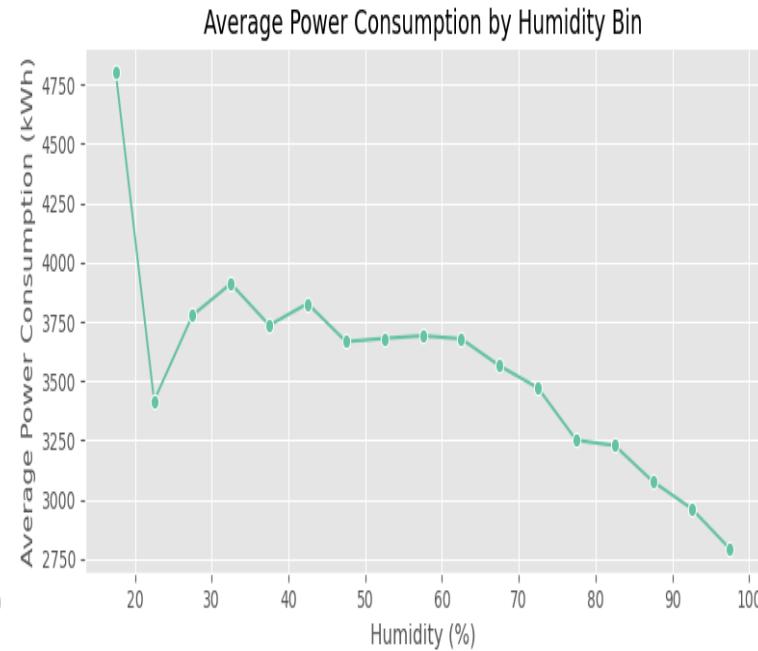
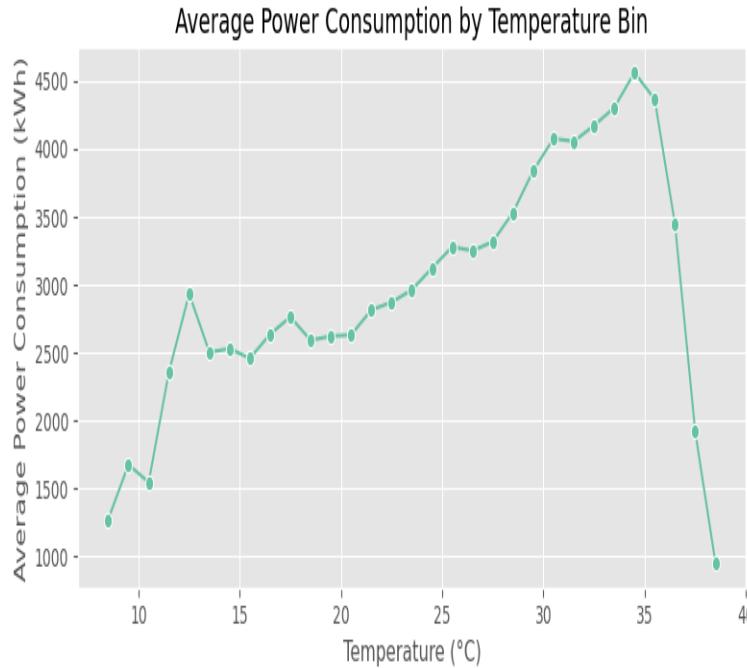


요일별 전력소비량 분석

평일에는 하루 평균 약 3,400kWh 수준으로 비교적 일정한 전력 소비 패턴
주말에는 평균 약 3,000kWh로 다소 감소하는 경향이 있다.

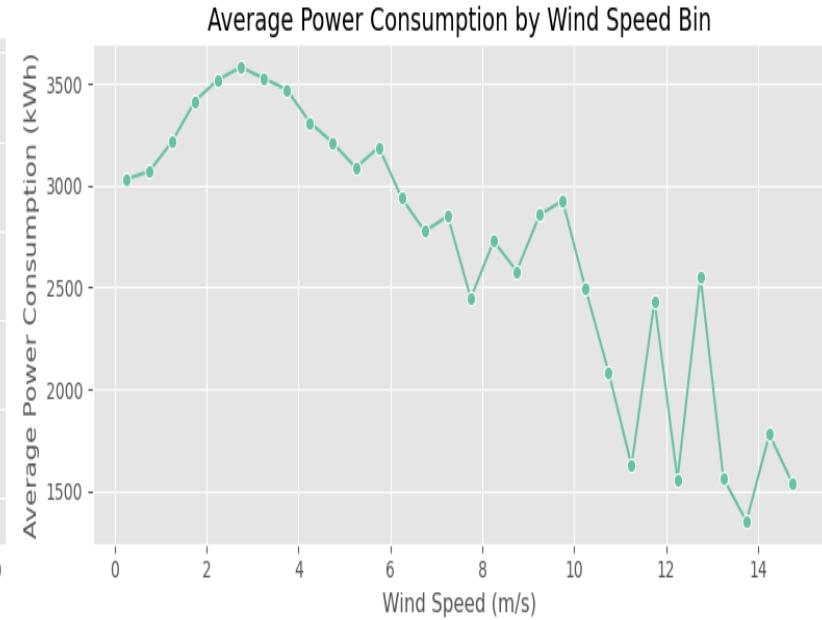
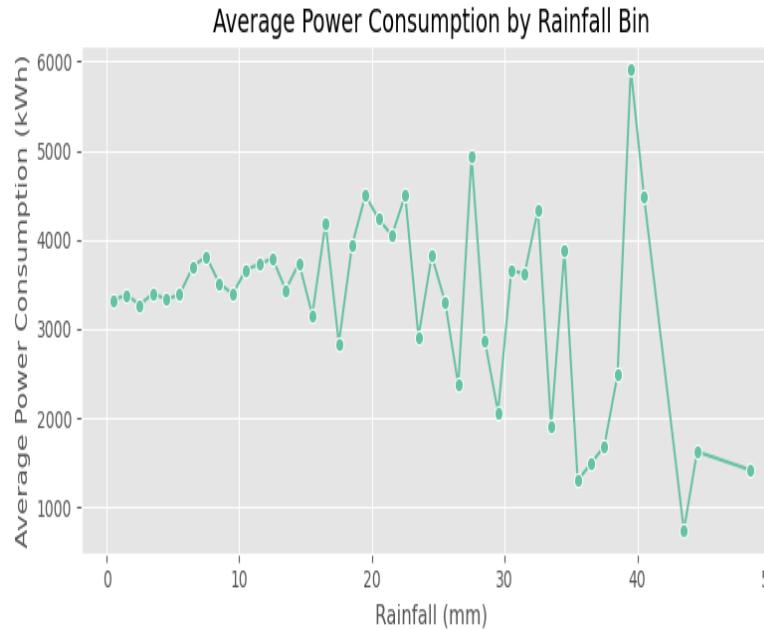


날씨 변수와 전력소비량 분석 (기온, 습도)



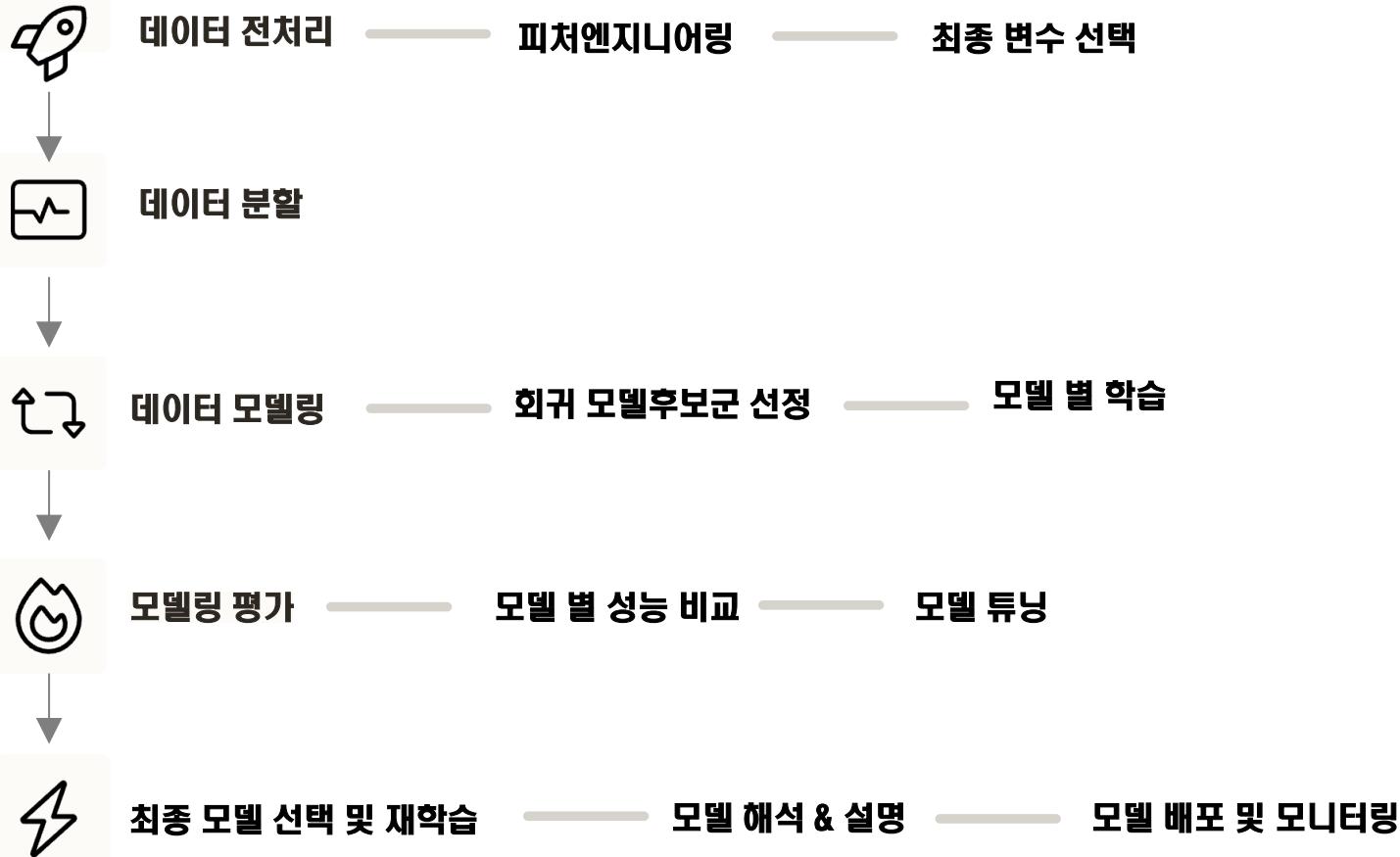
기온이 약 5°C 상승 할 때 전력사용량 또한 약 500kwh 상승하고
습도가 높아질수록 전력소비량은 점진적으로 감소하는 경향이 있다

날씨 변수와 전력소비량 분석 (강수량, 풍속)



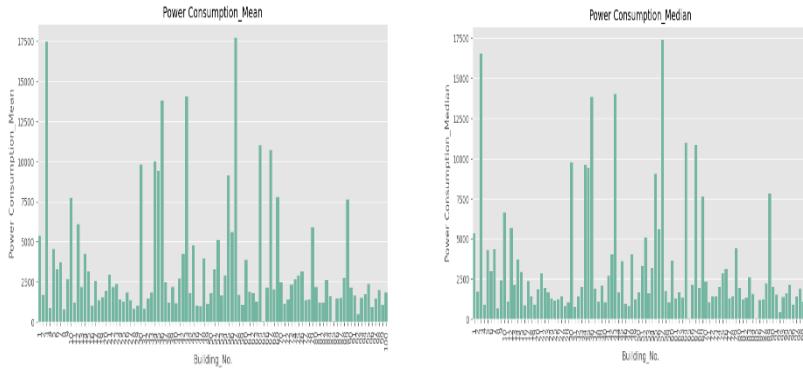
강수량과 풍속이 상승[증가]할 때 전력소비량은 감소하는 경향이 있다

ML 파이프 라인 구축



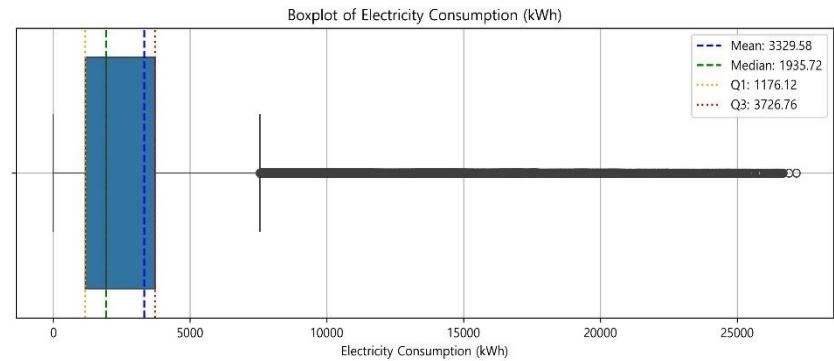
데이터 이상치

건물별/중앙값 비교



- ✓ 이상치가 평균에 영향을 주고 있음
- ✓ 일부 건물은 평균과 중앙값이 거의 유사 함.
- ✓ 건물 용도, 규모, 운영 시간등 특성차이
- ✓ 도메인 기반 이상치 해석
 - 대형 건물의 전력소비가 높은것은 당연함.
 - 에어컨, 조명 등 주기적인 소비
 - 건물 크기, 용도에 따라 전력소비량 상승

이상치 존재 여부



- ✓ Mean : 3,329kwh, Median : 1,935kwh
- ✓ Q1 = 1176kwh, Q3 = 3726kwh ,
- ✓ IQR = 2550kwh
- ✓ 이상치 상한선 $Q3 + 1.5 * IQR = 7552\text{kwh}$



이상치가 존재하나 도메인 기반 해석의 중요성으로 삭제하지 않음.

데이터 결측치

태양광 용량
0
278.58
0
1983.05

- ✓ ESS, PCS 변수는 결측치 비율 88%, 삭제 조치
- ✓ 태양광용량 변수 결측치는 0으로 대체

원 핫 인코딩

Num_data_time
전력소비량(Kwh)
일조(hr)
일사(mj/m2)

N개

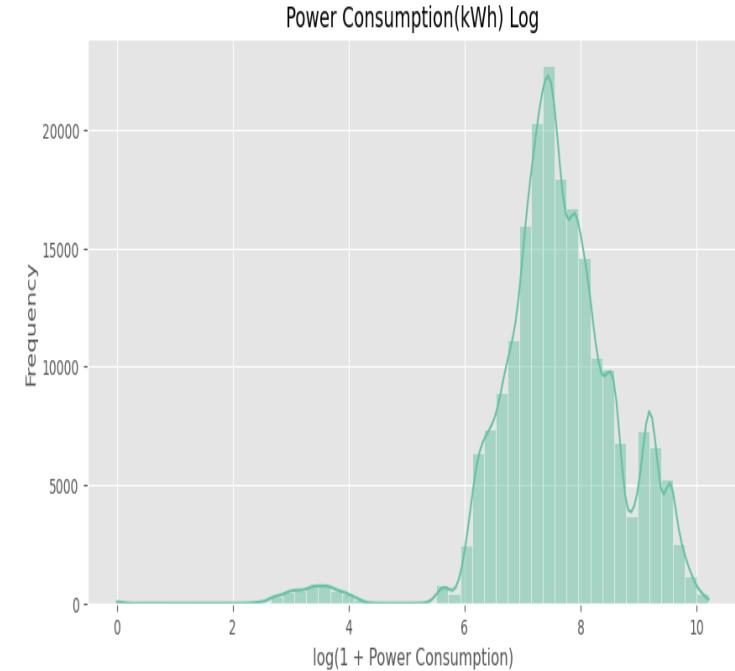
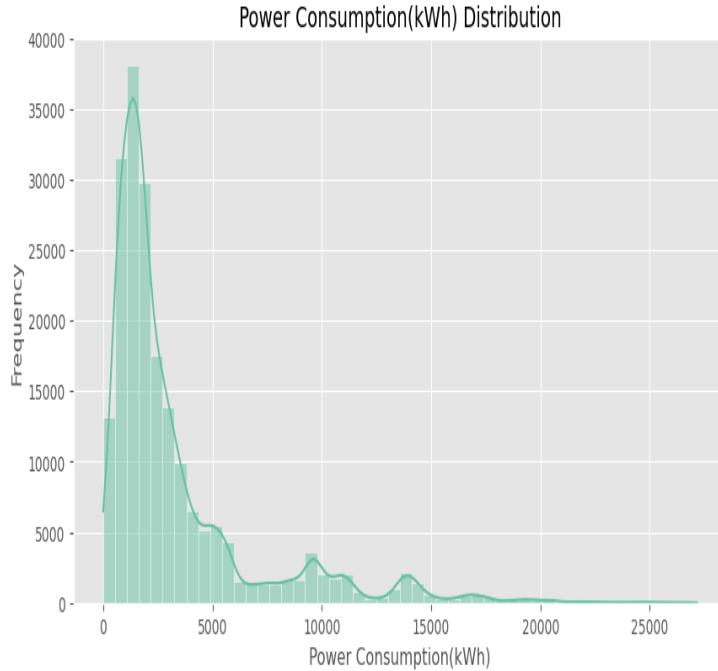


0	0	0
1	0	0
0	1	0
0	0	1

(N-1)개

- ✓ 범주형 변수를 다차원 백터로 변환
- ✓ 더미 변수제거를 통한 다중공선성 방지

로그 스케일링 전, 후 모델링 적용



0 ~ 5,000Kwh 밀집된 데이터를 log 스케일 기준 6 ~ 8 근처에 집중
3,000 ~ 4,000Kwh 극단값을 2,000 ~ 2,500kwh 까지 축소하여
데이터의 분포를 완화하여 모델이 더 쉽게 학습 가능하도록 노력함.

확정 변수

Building_Info.csv (5)	
건물번호	고유번호
건물유형	건물용도
연면적(m ²)	전체면적
냉방면적(m ²)	냉방가능 면적
태양광용량(KW)	태양광 발전용량
ESS저장용량	ESS용량
PCS용량	PCS용량

Train.csv (7)	
num_data_time	일시
건물번호	고유번호
기온	기온
강수량(mm)	강수량
풍속(m/s)	풍속
습도(%)	습도
전력소비량	예측대상
일조	일조량
일사	일사량
일시	일시

파생변수생성

파생변수명	설명
Month	월
Day	일
Hour	시간
Weekday	평일
Is_weekday	주말
df_train	범주형 인코딩

모델 실험 데이터화

Log 변환 전/후 모델링, 머신러닝 모델별 실험, 하이퍼파라미터 튜닝 등
 총 81번의 모델링 실험을 통해서 최적의 성능 높은 모델 선정

모델	MAE	RMSE	R2-Score	SMAPE
Logistic Regression	1742.4418	2617.2905	0.4922	61.1298 %
Ridge	1742.4419	2617.2911	0.4922	61.1291 %
Lasso	1741.9943	2617.3336	0.4922	61.0273 %
ElasticNet	2205.1553	3281.0518	0.2020	67.1693 %
DecisionTree	457.5209	744.1168	0.9590	20.3597 %
RandomForest	98.0914	212.3211	0.9967	4.3251 %
XGBoost	195.5019	320.3932	0.9924	12.2071 %
LightGBM	283.6727	432.3166	0.9861	16.2243 %

모델	MAE	RMSE	R2-Score	SMAPE
Logistic Regression	1742.4418	2617.2905	0.4922	61.1298 %
Ridge	1742.4419	2617.2911	0.4922	61.1291 %
Lasso	1741.9943	2617.3336	0.4922	61.0273 %
ElasticNet	2205.1553	3281.0518	0.2020	67.1693 %
DecisionTree	457.5209	744.1168	0.9590	20.3597 %
RandomForest	98.0914	212.3211	0.9967	4.3251 %
XGBoost	195.5019	320.3932	0.9924	12.2071 %
LightGBM	283.6727	432.3166	0.9861	16.2243 %

모델	MAE	RMSE	R2-Score	SMAPE
Logistic Regression	1736.5885	2612.4838	0.4941	61.0681 %
Ridge	1736.5891	2612.4844	0.4941	61.0675 %
Lasso	1736.0525	2612.5224	0.4941	60.9607 %
ElasticNet	2196.6101	3274.2603	0.2053	66.9431 %
DecisionTree	459.4153	754.0381	0.9579	20.6487 %
RandomForest	100.6937	221.1437	0.9964	4.4075 %
XGBoost	198.3500	327.3461	0.9921	12.4799 %
LightGBM	282.0174	429.6373	0.9863	15.8768 %

모델	MAE	RMSE	R2-Score	SMAPE
Logistic Regression	1736.2612. 5885	2612. 4838	0.4941	61.0681 %
Ridge	1736.2612. 5891	2612. 4844	0.4941	61.0675 %
Lasso	1736.2612. 0525	2612. 5224	0.4941	60.9607 %
ElasticNet	2196.3274. 6101	3274. 2603	0.2053	66.9431 %
DecisionTree	459.4754. 153	754.0381	0.9579	20.6487 %
RandomForest	100.6221. 937	221.1437	0.9964	4.4075 %
XGBoost	198.3327. 500	327.3461	0.9921	12.4799 %
LightGBM	282.0429. 174	429.6373	0.9863	15.8768 %

모델	MAE	RMSE	R2-Score	SMAPE
Logistic Regression	1648.4235	2777. 3857	0.4282	55.2414 %
Ridge	1648.4346	2777. 4160	0.4282	55.2415 %
Lasso	2187.0058	3891. 9280	-0.1228	65.4627 %
ElasticNet	2187.0058	3891. 9280	-0.1228	65.4627 %
DecisionTree	460.1511	834.0966	0.9484	18.1996 %
RandomForest	101.0576	222.3538	0.9963	4.4591 %
XGBoost	228.0133	407.7607	0.9877	8.7143 %
LightGBM	365.2502	670.7128	0.9667	12.6837 %

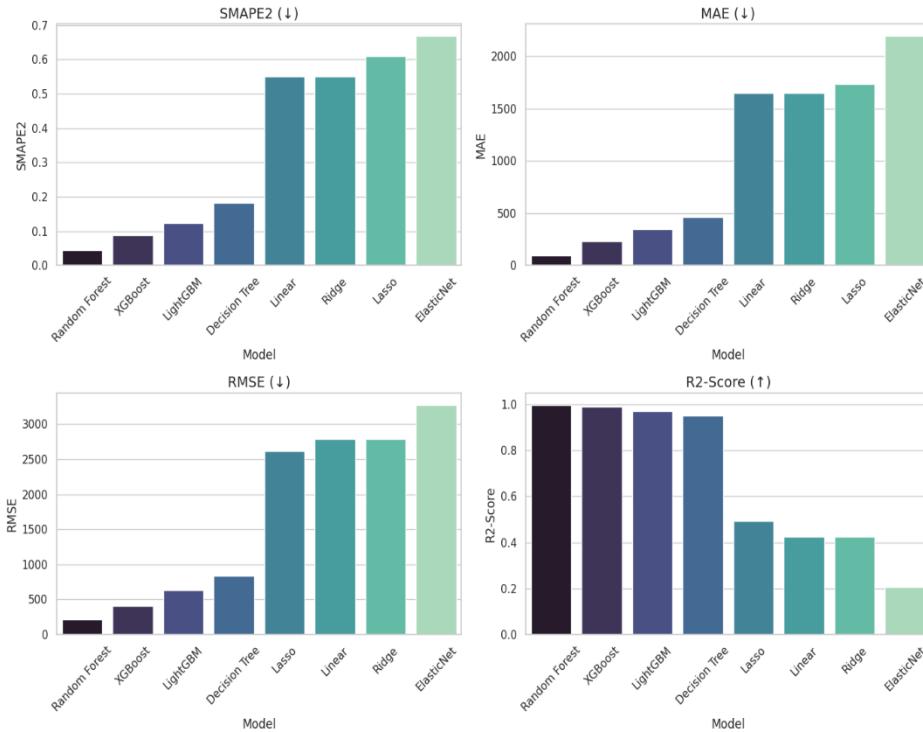
모델	MAE	RMSE	R2-Score	SMAPE
Logistic Regression	1648.4235	2777. 3857	0.4282	55.2414 %
Ridge	1648.4346	2777. 4160	0.4282	55.2415 %
Lasso	2187.0058	3891. 9280	-0.1228	65.4627 %
ElasticNet	2187.0058	3891. 9280	-0.1228	65.4627 %
DecisionTree	460.1511	834.0966	0.9484	18.1996 %
RandomForest	101.0576	222.3538	0.9963	4.4591 %
XGBoost	228.0133	407.7607	0.9877	8.7143 %
LightGBM	365.2502	670.7128	0.9667	12.6837 %

모델	MAE	RMSE	R2-Score	SMAPE
Logistic Regression	1649.7233	2784.1207	0.4254	55.1080 %
Ridge	1649.7346	2784.1484	0.4254	55.1081 %
Lasso	2187.0058	3891.9280	-0.1228	65.4627 %
ElasticNet	2187.0058	3891.9280	-0.1228	65.4627 %
DecisionTree	585.0211	1279.3715	0.8787	22.8537 %
RandomForest	105.5552	235.8484	0.9959	4.6314 %
XGBoost	234.6040	428.2717	0.9864	8.8036 %
LightGBM	344.6426	630.5697	0.9705	12.2333 %

최종 모델 선정

Random Forest

트리 기반 머신 러닝 모델 성능 우수

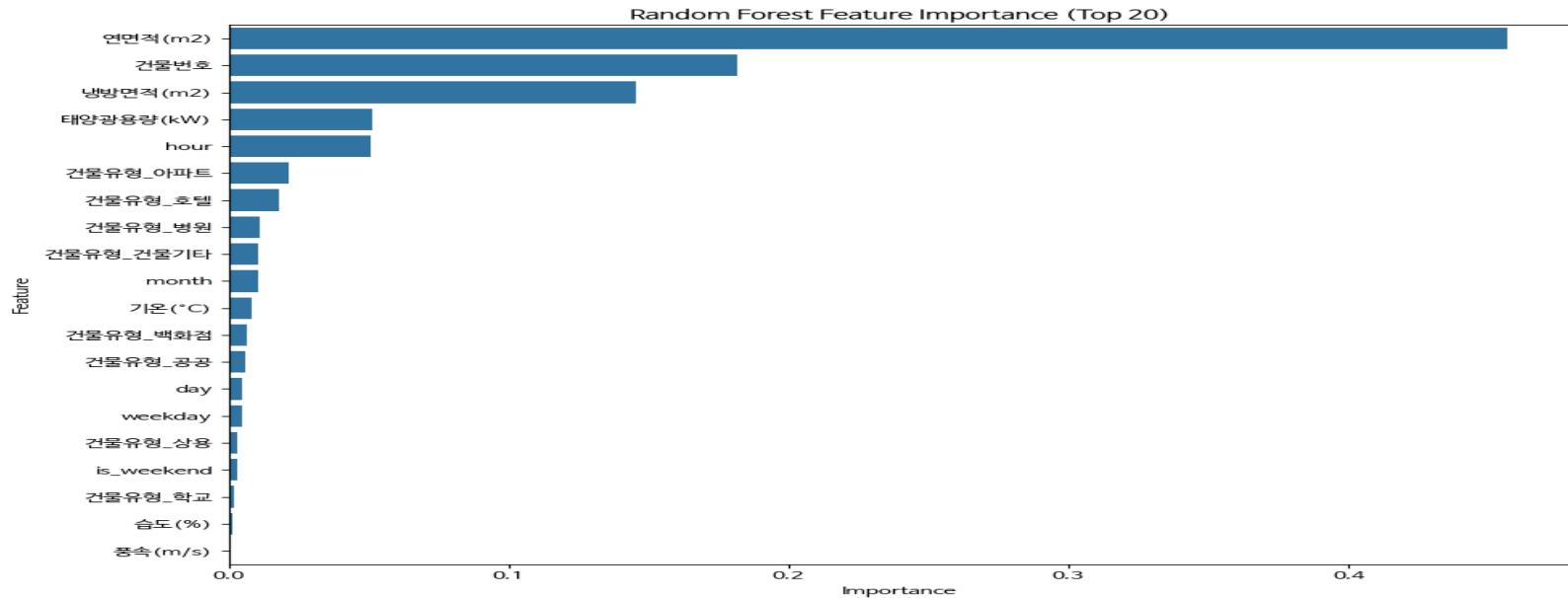


모델	MAE	RMSE	R2-Score	SMAPE
Logistic Regression	1742.4	2617.29	0.4922	61.1298 %
Ridge	1742.4	2617.29	0.4922	61.1291 %
Lasso	1741.9	2617.33	0.4922	61.0273 %
ElasticNet	2205.1	3281.05	0.2020	67.1693 %
DecisionTree	457.52	744.11	0.9590	20.3597 %
RandomForest	98.09	212.321	0.9967	4.3251 %
XGBoost	195.50	320.393	0.9924	12.2071 %
LightGBM	283.67	432.316	0.9861	16.2243 %



모델 데이콘 제출 결과

모델 기반 변수 중요도 Feature Importances



모델 변수 중요도를 분석한 결과

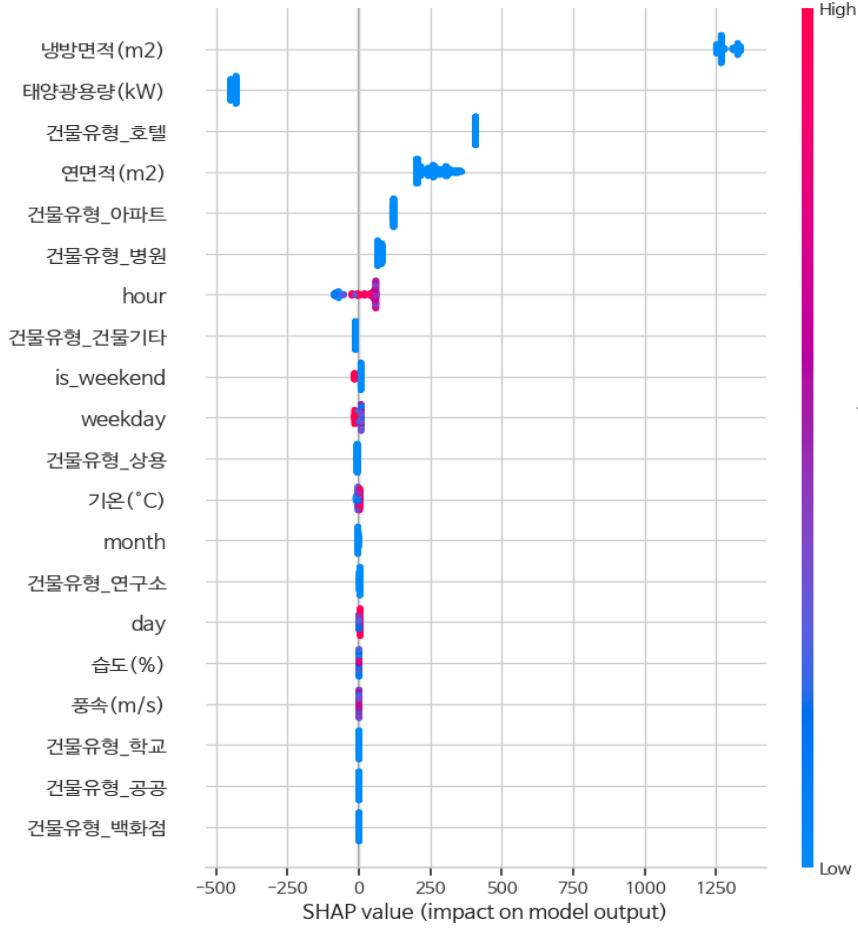
“연면적, 건물유형, 냉방면적, 태양광용량, 시간 등이 높은 중요도를 가졌음을 확인.



이 방법은 빠르게 주요 변수를 확인할 수 있다는 장점이지만
변수가 예측값에 미치는 실제 영향력은 잘 반영되지 않는다고 생각

SHAP 기반 중요도(Shapley Values)

전체 변수 SHAP



SHAP 기반 중요도 해석

각 입력값을 모델 예측에 기여한 정도를 개별

샘플 기준으로 계산

“냉방면적, 태양광용량, 건물유형, 연면적” 등이 큰 영향

단순히 중요도와 같이 많이 사용된 피처가 아니라,

피처의 영향력 순위



냉방면적은 전력 소비량에 가진 큰 영향

태양광용량은 전력 소비를 줄이는 데 기여하는 음의 요인

건물유형에 따라 시간, 요일, 주말도 중요한 소비 패턴

결정요인임.

기대효과

- ✓ 정확한 전력 수요 예측으로 안정적인 에너지 공급
 - 여름철 고온으로 인한 냉방 수요 급증시 사전 예측을 통해 예비전력 배치
- ✓ 에너지 효율성 향상 및 비용 절감
 - 건물 특성, 기온 변화, 태양광 용량을 통한 예측으로 운영비용 절감
- ✓ 스마트 그리드 및 사용자 맞춤형 에너지 관리 서비스 제공
 - 실시간 전력수요와 공급데이터를 바탕으로 에너지 절약 솔루션 제공 가능

팀원별 프로젝트 회고 (1/2)

★ 권태성 팀원 (데이터 전처리 및 EDA, 파이프라인 구축)

미니 팀 프로젝트 활동을 하면서 각자 다양한 전처리 아이디어를 제시해 보고, 가장 합리적이고 점수가 잘 나오는 방법으로 통일한 뒤 여러 모델로 실험해 결과를 비교했습니다. 처음엔 많이 막막했지만 프로젝트를 직접 경험하면서 어떤 부분이 부족한지 명확히 파악할 수 있었고, 이 과정을 통해 실력이 크게 향상되었다고 느꼈습니다. 다음 프로젝트에서는 팀원들에게 더 유익한 도움을 줄 수 있도록 준비를 더욱 철저하게 해야겠다는 다짐을 하게 되었습니다.

★ 한대유 팀원 (데이터 전처리, 모델링 전문, 모델링 성능 실험)

이번 프로젝트를 통해 EDA, 데이터 전처리, 모델링을 하면서, 데이터 전처리가 모델 성능에 얼마나 영향을 주는지를 체감할 수 있었습니다. 모델의 예측 성능을 올리기 위해 여러 차례 하이퍼 파라미터 튜닝을 했으나 성능을 개선하지 못하고 랜덤 포레스트의 기본 설정 모델이 가장 좋은 성능을 보인 부분은 조금 더 좋은 성능을 낼 수 있지 않았을까 하는 아쉬움이 남았습니다. 다음 프로젝트에는 하이퍼 파라미터 튜닝을 통해 모델 성능을 개선하고, 양상을 기법을 활용하여 더 좋은 결과를 얻을 수 있었으면 좋겠습니다.

팀원별 프로젝트 회고 (2/2)

★ 정안식 팀원 (데이터 전처리 및 EDA, 파이프라인 구축)

이번 프로젝트에서는 건물 정보와 기상 데이터를 통합하고, 월, 일, 시간, 요일, 주말 등의 파생 변수를 생성하여 시계열 특성을 반영한 전처리를 수행했습니다. 특히 test 데이터에 누락된 일조/일사 칼럼은 별도 회귀 모델을 통해 예측값을 생성해 삽입함으로써 모델 입력 특성의 일관성을 유지하고자 했습니다. 또한 ‘-’ 문자 처리, 이상치 제거, 스케일링, 인코딩의 기본 전처리를 적용했습니다. 데이터 품질이 모델 성능에 영향을 준다는 점을 실감할 수 있었던 경험이었습니다.

★ 김도현 팀원 (문제정의, 모델링 주도, 프로젝트 리딩)

프로젝트 일정을 조율하고 흐름을 정리하는 역할을 하며, 팀원 각자의 전문성과 협업의 시너지를 느낄 수 있었습니다. 데이터의 전처리에 많은 시간을 공들였는데 모델링 과정에 이러한 점이 많이 반영된 거 같습니다. 또한 모든 결과를 시각화를 통해 팀원들과 인사이트를 공유하고 전체적인 흐름을 파악하는 데 큰 도움을 줄 수 있었던 것 같습니다. 다음에는 아이디어를 발전시켜 사회 문제에 직접적인 도움을 줄 수 있는 데이터를 통해서 조금 더 의미 있는 프로젝트를 해서 사회문제에 적용해 볼 수 있지 않을까 생각했습니다.