시공간 복합데이터를 활용한 전력 수요예측 개선

참가번호 240555 팀명 카이제곱분포

1. 연구 배경 및 필요성

전 세계적으로 전력 수요량은 2000년대 이후로 꾸준히 증가하는 추세이다. 그 가운데 한국의 전력 소비량 증가는 눈에 띄게 빠른 수준이다. 2023년 국제 에너지기구(IEA)의 발표에 따르면, 2021년 기준 대한민국은 세계에서 3번째로 전력 소비량이 큰 나라로 조사되었으며, 이는 한국의 인구 규모를 고려했을 때 한국의 인당 전력 소비량이 타국에 비해 얼마나 큰 수준인지 가늠케 한다. 이처럼 규모의 전력을 소비하고 유통하는 한국에서 전력 수요의 예측이란 필수적인 문제이다. 여름철 급격히 증가하는 전력 수요에 유연하게 대응하고, 적절한 수준의 전기료를 책정하기 위해서는 사전에 전력 수요량을 예측할 수 있어야 한다.

본 분석에서는 기본적으로 제공된 기온, 상대습도와 같은 시계열 기상 관측 변수 데이터와 더불어, 각 관측 지점의 공간적인 위치 자체를 변수로 사용하여 시공간적인 특성에 기반한 전력 수요를 예측할 수 있도록 모델링하고 각 변수의 영향력을 검토하였다.

2. 활용 데이터셋

※ 파생변수는 파란색으로 하이라이트 함.

아래 표 (1)은 본 공모전에서 기본적으로 제공된 기상 전력 학습 데이터 및 그로부터 생성한 파생변수들의 일람이며, 표 (2)는 표(1)의 STN변수와 관련지어 관측지점의 지리적인 정보를 반영할 수 있도록 한 관측지점 정보 데이터 및 그의 파생변수 일람이다.

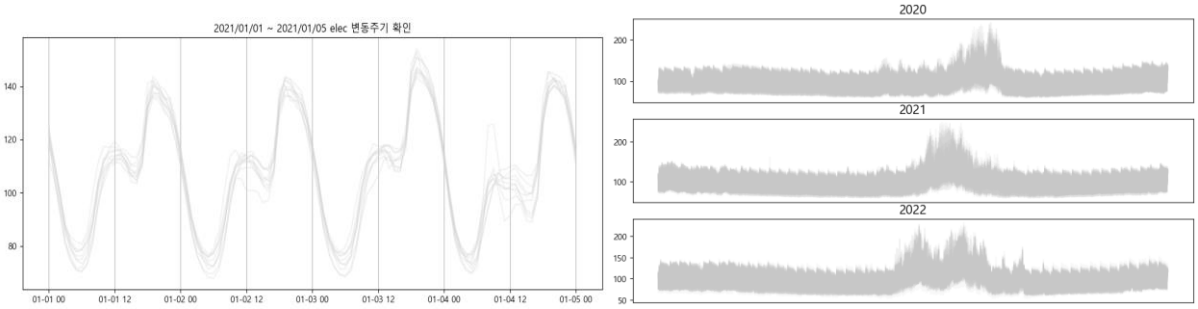
(1) 기상 전력 학습 데이터 (시계열 데이터)

Column	Description	Column	Description
NUM	기상 관측 격자 번호	afternoon	12-15시를 1, 나머지를 0 표기
STN	AWS 지점 번호	nph_ta	기온 (℃)
N	격자 내 공동주택의 수 (단지)	nph_hm	상대습도 (%)
TM	전력부하 측정날짜 (시간 단위)	nph_ws_10m	10분 평균 풍속 (m/s)
HH24	전력부하 측정시간 (1~24)	nph_rn_60m	1시간 누적 강수량, 단위(mm)
weekday	요일 (0~6)	nph_ta_chi	체감온도, 단위(℃)
week_name	주중/주말 (0/1)	denoise_ta	Nph_ta를 푸리에 변환
sin_time	HH24 변수에 sin으로 주기성 반영	denoise_hm	Nph_hm를 푸리에 변환
cos_time	HH24 변수에 cos로 주기성 반영	denoise_ws	Nph_ws_10m를 푸리에 변환
sin_month	TM의 월 값에 sin으로 주기성 반영	denoise_ta_chi	Nph_ta_chi를 푸리에 변환
cos_month	TM의 월 값에 cos로 주기성 반영	THI	불쾌지수
summer_sin	여름 특징 반영 위해 sin 변환	불쾌여부	불쾌여부 (0/1)
summer_cos	여름 특징 반영 위해 cos 변환	elec	전력기상지수

(2) ★ 관측지점 정보 데이터 (공간 데이터)

Column	Description	Column	Description
AWS	기상 전력 학습 데이터의 STN에 대응하는 변수	고도	AWS 관측지점의 고도 (m)
경도	AWS 관측지점의 경도	location	AWS 관측지점의 행정구역
위도	AWS 관측지점의 위도	mountain	AWS 관측지점의 산지 입지 여부
수도권 여부	AWS 관측지점의 수도권 여부		

주기성을 반영하는 파생변수 생성



▲ (左) 일별 주기를 나타낸 시각화, (右) 연도별 주기를 나타낸 시각화.

왼쪽의 시각화는 주어진 학습 데이터셋의 격자 중 랜덤하게 15개를 추출하여 2021/01/01 ~ 2021/01/05 에 해당하는 5일간의 elec 변수의 변동을 겹쳐 보인 결과다. 매일 0시를 기준으로 동일한 개형이 반복되는 주기성을 가지고 있음을 확인할 수 있다. 따라서 해당 변수가 가진 계절성을 Fourier Features에 기반해 삼각함수로 변환함으로써 (sin, cos) 주기적 특성을 더욱 적극적으로 반영할 수 있도록 한다. (sin_time, cos_time)

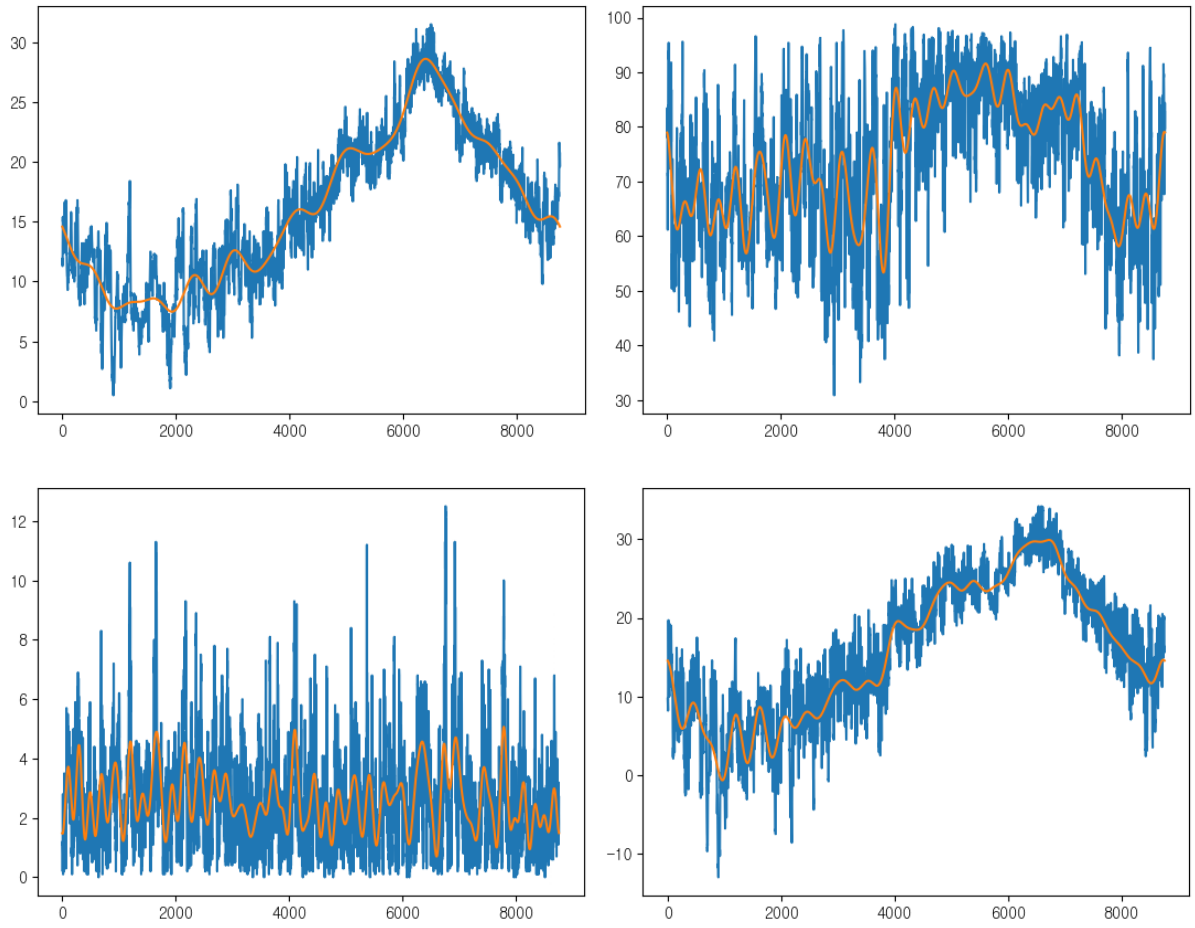
오른쪽의 시각화는 주어진 학습 데이터셋의 격자 중 랜덤하게 10개를 추출하여 2020 ~ 2022년에 해당하는 3개년의 elec 변수의 변동을 연 단위로 겹쳐 보인 결과다. 연 주기로 elec 변수가 유사한 개형을 가지고 변동하고 있음을 확인할 수 있다. 위와 동일한 논리로 1~12 주기의 TM으로부터 추출된 월 변수를 삼각함수를 이용해 변환하였다. (sin_month, cos_month)

마지막으로, 오른쪽 시각화에서도 확인할 수 있듯이 여름(6~8월)에 해당하는 elec 변수의 변동이 다른 월별 주기와 구분되는 특징을 가지고 있다는 점을 반영하기 위해 여름 자체의 주기를 반영할 수 있도록 삼각함수를 이용해 주기성을 띄도록 변환 하였다. (summer_sin, summer_cos)

afternoon 변수 역시 삼각함수를 사용하지는 않았지만 오전/오후라는 주기를 나타내는 파생변수로서 생성하였다. 왼쪽의 시각화에서도 확인 가능하듯, 오후 12시 이후를 기준으로 elec 변수가 NUM마다 편차가 크게 변동하고 있음을 고려하였다. 따라서 NUM간 분산이 큰 12시~15시 구간에 해당하는 데이터를 1, 그 외의 데이터를 0으로 표기하는 이진변수를 생성하였다.

푸리에 변환을 적용한 파생변수 생성

노이즈가 많이 끼어있어 변동이 불안정한 원 시계열 변수들에 대해 푸리에 변환을 적용하여 실제 기상 상황의 변동을 정확하게 반영하고자 하였다. 아래 4개의 시각화는 차례대로 'nph_ta', 'nph_hm', 'nph_ws_10m', 'nph_ta_chi' 시계열 기상변수들에 대해 푸리에변환을 적용하기 전(파란색)과 후(주황색)의 비교를 나타낸 결과다.



▲ 차례대로 'nph_ta', 'nph_hm', 'nph_ws_10m', 'nph_ta_chi' 의 푸리에 변환 전/후.

노이즈를 제거하여 원 시계열 데이터의 변동 파형을 추출하였으며(주황색) 이를 새로운 파생 변수로 삽입하였다.

불쾌지수 및 불쾌 여부 파생변수 생성

전력사용량이 여름철 급격하게 변동함을 적극적으로 반영하기 위하여 불쾌지수를 계산하여 파생변수로서 생성하였다. 불쾌지수는 기온과 습도를 이용하여 계산되는 값으로, 그 공식은 아래와 같다.

$$\text{불쾌지수} = \frac{9}{5}\text{기온} - 0.55(1 - \text{습도})\left(\frac{9}{5}\text{기온} - 26\right) + 32$$

위의 공식으로 산출된 불쾌지수가 약 80을 초과하는 경우 큰 불쾌감을 초래한다는 국민건강보험의 의견에 기반하여 80이상의 불쾌지수를 산출한 데이터를 1로, 그렇지 않은 데이터를 0으로 표기하는 이진변수를 추가적으로 생성하였다.

위도/경도/고도 변수 도입

시계열적인 특징 뿐만 아니라, 시계열 데이터의 관측 지점이 가지는 지리적인 특징이 유의미할 수 있다는 가설 하에 외부 데이터로부터 위도/경도/고도 변수를 도입하였다. 지리적인 특징은 에너지 소비에 있어 사회문화적으로 유의미한 영향으로 파생되어 드러날 수 있을 것이라고 판단하였기 때문이다. 실제로 국제 에너지기구(EIA)의 2015년 지역별 가

정용 전기 소비량 레포트에서는, 미국 남부지역이 평균적으로 연간 약 1.4만 킬로와트를 소비한 데에 반해 중서부 지역은 연간 9천 킬로와트를, 북동부지역은 8천 킬로와트를 소비했다는 것을 확인할 수 있었다.[i]

행정구역 변수 도입 (location)

AWS 관측지점의 도 단위 행정구역을 변수로 추가하였다. AWS 관측지점의 구체적인 지명 자체(군, 면, 리단위) 를 변수로 사용할 수도 있었으나, 여러 개의 고유한 값으로 구성된 문자열 변수를 머신러닝 모델에 학습하는 경우 변수 차원의 개수가 급격하게 증가(718차원 증가)하여 모델의 불안정성이 증가할 위험을 우려하였다. 따라서 안정적으로 모델을 학습시키면서도 관측지점의 지리적 특성을 반영할 수 있는 상위 계층의 행정구역(15차원)을 최종 추가 변수로 채택하였다.

mountain 변수와 수도권 여부 변수

AWS 관측지점의 지리적인 특징으로서 유의미할 것으로 추측되는 산지 입지 여부를 이진 변수로 추가하였다. 또한, 지역간 사회문화적인 이질성(인구밀도 등)이 가지는 영향력을 고려하여 수도권 여부를 이진변수로 추가하였다.

3. 모델링

기상에 따른 공동주택 전력 사용량을 분석하는 본 과제는 분석에 있어 다음과 같은 특징을 보이고 있다. (1) 지역과 시간이라는 두 차원이 혼재한다는 점과 (2) 반응변수인 전력 사용량이 연 평균을 기준으로 스케일링된 상대적 지표라는 점이다. 특징 (2)로 인해, 반응변수에서는 지역적 특성이 다소 약화되었으나 일부 설명변수는 그 값이 절대적 지표이기에, 지역적 특성이 반응변수에 비해 큰 것으로 판단되었다.

즉, 시간적 특성에 각 공간에 따른 차이가 혼재되어 있는 경시적 자료의 특성을 지니기에 본 모델에서 기존의 전통적인 시계열 모형을 사용하여 예측하는 것은 다소 무리가 따른다고 결론지었다. 이에, 대용량의 데이터에 강건한 tree 기반의 Custom 모델을 통해 데이터의 시계열적인 특성을 살리면서 분석의 정확도를 높이고자 했다.

데이터의 시계열적인 특성을 최대한 반영하기 위해 train set과 validation set은 무작위 추출이 아닌 2020~2021년 까지의 데이터는 train set으로, 22년 데이터는 validation set으로 구성하여 모델 평가에 활용하였다. train set과 validation set의 비율은 약 7:3 으로 구성되었다.

해당 Custom 모델은 데이터의 시계열적 요소를 파악하는 1차 모델과, 잔여 오차에 대해 추정하는 2차 모델로 구성되어 있다. 두 모델 모두 반응변수에 존재하는 노이즈로 인해 과적합의 우려가 있어 하이퍼 파라미터 튜닝은 따로 진행하지 않은 상태로 모델링이 진행되었다.

1차 모델은 반응 변수의 시간적 요소를 최소화하는 방향으로 모델링이 진행되었다. 이에, 데이터의 시간적 특성을 최대한 반영하고자 기존 기상 변수 중 시계열적 특성을 보이는 변수와, 주기성을 보이는 파생변수 등을 활용하여 모델링을 진행할 데이터셋을 구성하였다. 1차 모델에 사용된 변수는 아래와 같다.

nph_ta	nph_ta_chi	denoise_ws	cos_time
nph_hm	denoise_ta	denoise_ta_chi	sin_month
nph_ws_10m	denoise_hm	sin_time	cos_month

THI			

다만, 전통적인 선형 시계열 모델의 경우 해당 시계열이 비정상성 시계열의 모습을 보이며 여름에 그 수치가 급격하게 증가하는 비선형 추세를 보이기에 모형을 적합하기 어렵다고 판단하였다. 또한 LSTM 등의 딥러닝 시계열 모형을 활용하는 것 역시 설명변수와 반응변수에 존재하는 노이즈로 인해 과적합의 우려가 있다고 판단하였다. 또한 결정적으로, 시계열적 특성을 다루는 모형에서는 공간적 특성을 반영할 수 없기에, 시계열적 모형이 아닌 머신러닝 모형을 통해 시간적 특성을 파악하고자 하였다.

사용된 모형은 LGBM으로, LGBM은 대용량 데이터셋에 강건하며 데이터의 비선형적 특성을 파악하는데 강력한 성능을 보인다는 특징이 있다. 이를 통해 데이터의 시간적 특성을 파악한 결과는 아래와 같다.

Metric	MSE	MAE	R^2
Value	62.43	5.80	0.906

2차 모델은 1차 모델의 잔차에 대한 학습을 통해 진행된다. 해당 데이터에서는 시계열적 특성이 대부분 약화되었기에, 1차 모델링에 사용된 일부 시계열 변수와 2차 모델링에 사용되지 않았던 기타 파생변수들을 통한 모델링을 진행하였다. 이때, 일부 시계열 변수가 사용된 이유는 오차의 특정 시간대의 예측력이 떨어지는 등 오차에 일부 시계열적 특성이 반영되었기 때문이다. 이에 사용된 변수는 아래와 같다.

HH24	Weekday	경도	Cos_time
nph_ta	Week_name	위도	Afternoon
nph_hm	Year	고도	Summer_cos
nph_ws_10m	Month	mountain	Summer_sin
nph_ws_60m	Day	수도권여부	불쾌여부
nph_ta_chi	summer	Sin_time	

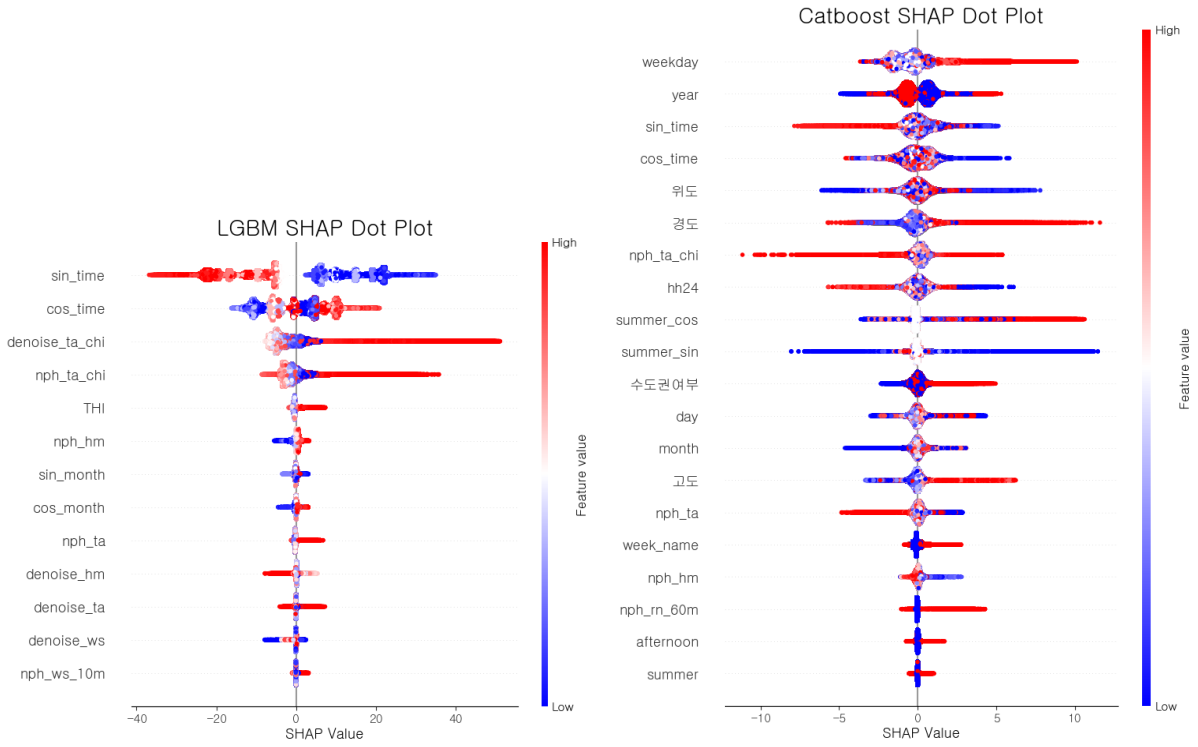
해당 데이터는 범주형 자료가 다수 포함되어 있으며, 2차 모델의 경우 반응변수가 오차이기에 그 값에 대한 예측력이 떨어진다면 전체 전력 사용량의 예측 성능을 크게 저하시킬 수 있다는 문제점을 내포하고 있었다.

이에, 2차 모델에는 범주형 자료의 처리에 강력한 성능을 보장하고 Gradient Boosting 기반 모델 중 과적합에 강건한 Catboost Regressor를 사용하여 1차 모델에 대한 잔차를 보정하였다. 잔차에 대한 보정이 완료된 모델의 최종 예측 결과는 아래와 같이 나타났다.

Metric	MSE	MAE	R^2	ELEC_AVG
Value	39.33 ▼	4.37 ▼	0.940 ▲	**0.978**

전반적으로 LGBM 단일 모델링에 비해 MSE가 약 23.1 감소하고, $\boldsymbol{R^2}$가 0.35 증가하며 전반적인 예측 성능이 향상된 것을 확인할 수 있었다.

4. 모델링 결과 해석



▲ (左) LGBM 1차모델의 SHAP Plot, (右) CatBoost 2차 잔차 모델의 SHAP Plot

두 모델의 학습 결과를 기반으로 생성된 SHAP Plot으로 각 변수가 가지는 의미를 해석한다. 먼저 왼쪽의 LGBM SHAP Plot에서는 sin_time의 분리가 두드러진다. sin_time의 값은 오전에 양수, 오후에 음수 값을 가지므로 낮부터 오후까지에 해당하는 시간대에 전력 소비량이 극대화됨을 의미한다. 더불어 체감온도를 나타내는 nph_ta_chi 변수의 수치가 높을수록 전력 소비량은 양의 방향으로 늘어나지만, 그에 비해 체감온도가 낮아진다고해서 전력 소비량이 감소하지는 않는 비대칭성을 확인할 수 있다.

1차 LGBM 모델의 잔차에 대해 모델링한 CatBoost 2차 모델에서는 weekday 변수의 SHAP Value가 두드러진다. weekday의 값은 월~일에 각각 0~6을 대응시킨 변수이므로, weekday와 잔차가 양의 관계를 가진다는 것은 주말(5~6)에 전력 소비량 예측의 오차가 모델에 의해 크게 계산되었음을 의미한다. 이는 주말 시간대 주거지역에서 시간을 보내는 인구의 증가가 영향을 끼친 것으로 추측 가능하다. 또한 sin_time은 잔차에 대해 강력한 음의 영향력을 가지는 것으로 보이는데, 이는 왼쪽의 LGBM SHAP Plot에서 확인한 것과 일맥상통한다.

5. 의의

기상 시계열 데이터 및 지리데이터를 이용해 전력 수요량 예측을 모델링하고, 다양한 평가 지표를 통해 예측의 실현 가능성을 점검하였다. 본 분석에서는 주어진 기상 시계열 데이터와 기본적인 지리 데이터만을 사용하여 모델링하였지만, 추가적인 사회문화적인 변수의 도입을 통한 새로운 인사이트의 발굴 가능성을 검토해볼 수 있을 것으로 기대해볼 수 있다.

---

[i] U.S. Energy Information Administration, 2015 Residental Energy Consumption Survey