산업통상자원부 공공데이터 활용

비즈니스아이디어 공모전 분석 결과 제출

소속: 고려대학교

팀원: 노연수, 민윤기, 임채명

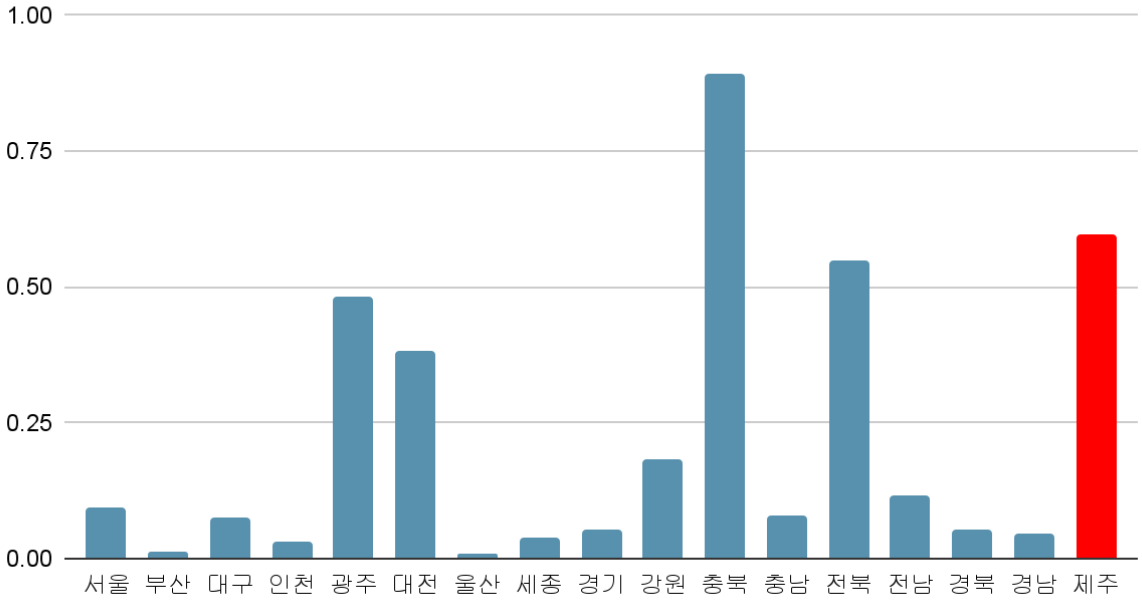
Ⅰ. 명칭

제주 지역 전력사용량 예측 모델 구축

Ⅱ. 제안배경

전력계통을 운영함에 있어 중요한 목표는 안정적이면서도 경제적인 전력 공급을 위한 정확한 전력수요 예측이다. 만일 수요예측에 있어 큰 오차가 생긴다면, 블랙아웃이나 2011년 국내에서 발생했던 대규모 순환단전과 같은 문제가 발생하게 된다. 따라서 지역을 막론하고서 안정적인 전력 수급을 위해 전력수요 예측은 꼭 필요하다. 특히 제주도의 경우, 독립적인 전력 예측이 필요하다. 그 이유는 다음과 같다.

1. 높은 신재생 에너지 보급률과 이에 따른 전력 불균형 현상: 제주특별자치도는 몇 년 전, '카본 프리 아일랜드(CFI)' 정책을 통한 신재생에너지 보급을 확대한 바 있다. 올해 5월 한국전력에서 발행한 공공데이터 '한국전력통계(92호)'에 제시된 '행정구역별 발전설비 및 발전량' 통계자료에 따르면, 2022년 제주 지역의 신재생에너지 발전량은 총 발전량의 60% 수준으로, 충북을 제외한 전국 17개 시도를 통틀어서 압도적으로 높은 신재생에너지 발전 비율을 보이고 있다. 신재생에너지는 환경친화적이면서 장기적인 비용 절감에도 도움이 되는 미래에너지로 주목받고 있지만, 최근 들어 제주 지역에서는 태양광 설비가 급증하는 등 신재생에너지 보급 확대에 따른 전력 불균형 현상이 두드러지고 있다. 이에 전기 수요가 공급 능력을 초과하거나, 반대로 전기가 과잉 공급됨에 따른 대규모 정전이 우려되는 상황이다.



2. 육지와는 다른 기후: 대부분의 신재생에너지 발전은 기후의 영향을 크게 받는다고 알려져 있다. 다시 말해, 제주 지역에서 신재생에너지 발전 비율이 높은 만큼, 전체적인 전력 수급의 불확실성은 클 수 밖에 없다. 또한, 전국(육지부)과 달리 산업단지나 대규모 기업 등이 없는 제주 지역의 전력수요 피크시간대는 연중 18~21시에 걸쳐 분포한다. 반면 전국의 경우, 여름과 겨울 각각 10~11시와 17시를 전후로 하여 전력수요 피크시간대가 형성되곤 한다.

위와 같은 이유로, 제주 지역은 타 도시와는 다르게 독립적으로 전력수요 예측이 진행되어야 하며, 예측된 전력수요량을 바탕으로 필요한 만큼의 에너지를 생산하는 것이 중요하다. 따라서 본 데이터 분석을 통해, 기상 조건변화에 즉각적으로 대응할 수 있는 제주 맞춤형 전력수요 예측을 목표로 예측 모델을 구축하고자 한다.

본 분석에서는 각 고객의 전력 사용에 영향을 미치는 기상 요소들을 주된 예측변수로 보고, 전력거래소에서 제공하는 하루전날 전력수요량 예측값을 포함하여 과거 코로나19 유행 기간의 사회적거리두기 단계 등을 고려하여 앞으로의 전력사용량을 예측한다. 나아가, 각 주택 단위에서의 예측을 뛰어넘어 제주 지역 전체의 전력수요 예측 모델로 발전시켜, 기상 조건을 포함한 다양한 사회적 환경에 따른 전력 수급의 안정성을 유지하는데 해당 분석 알고리즘을 제안하고자 한다. 아울러, 효율적인 신재생에너지 발전 및 적절한 양의 예비 전력 확보를 위한 중요한 참고 기술로 활용됨을 목표로 본 데이터 분석을 구상하였다.

Ⅲ. 분석 내용 및 분석 결과

3-1) 변수 설명

client_num: 고객번호

temp: 기온

wind: 풍속

humid: 습도

discomfort: 불쾌지수

temp_body: 체감온도

sin_hour: 사인함수 적용 시간

cos_hour: 코사인함수 적용 시간

weekday: 요일(0=월요일, 1=화요일, … , 5=토요일, 6=일요일)

day: 일

month: 월

year: 연도

weekend: 주말 여부(0=평일, 1=주말)

holiday: 평일 공휴일 여부(0=평일 공휴일 아님, 1=평일 공휴일)

dist0: 거리두기_해제(1=거리두기 전면 해제)

dist1.5: 거리두기_1.5 단계(1=1.5 단계 시행)

dist2: 거리두기_2 단계(1=2 단계 시행)

dist3: 거리두기_3 단계(1=3 단계 시행)

dist4: 거리두기_4 단계(1=4 단계 시행)

dist_with: 거리두기_위드코로나(1=위드코로나 거리두기 시행)

dist_with_up: 거리두기_위드코로나 강화(1=위드코로나 강화된 거리두기 시행)

dist_down1: 거리두기_위드코로나 1 차 완화(1=위드코로나 1 차 완화된 거리두기 시행)

dist_down2: 거리두기_위드코로나 2 차 완화(1=위드코로나 2 차 완화된 거리두기 시행)

dist_down3: 거리두기_위드코로나 3 차 완화(1=위드코로나 3 차 완화된 거리두기 시행)

cluster: 고객별 군집화 결과

client_mean: 고객별 전력사용량 평균

standard_prediction: 표준화된 하루전 발전계획용 전력 수요 예측량

target: 전력사용량(y)

3-2) 예측변수 설명

전력 사용량을 예측할 때 고려한 요인으로 기상 변수, 시간 변수, 코로나 19 관련 변수, 하루전 발전계획용 전력 수요예측량이 있다.

첫째, 기상 변수로 기상청 데이터의 기온, 풍속, 습도와 더불어 기온과 습도가 고려된 불쾌지수, 기온과 풍속이 고려된 체감온도를 계산하여 활용했다.

기온의 경우 여름철 폭염과 습한 날씨로 냉방 기구 사용이 증가하고, 겨울철 한파와 폭설로 난방 기구 사용량이 증가함에 따라 전력 사용량도 증가하게 된다. 기온은 전력 사용량과 밀접한 관련이 있으며, 여러 선행 연구들에서 기온이 전력 수요에 큰 영향을 미친다고 확인되었다. 김혜민 등 (2015)의 "기온이 전력수요에 미치는 영향 분석"에 따르면 전력수요함수를 추정함에 있어서 기온 변수를 추가하는 것이 그렇지 않은 것보다 통계적으로 유의한 결과를 보였다.

전력수요를 예측할 때 많이 활용하는 기상요인은 기온인데, 기상청에서 보도된 자료에 따르면 기상 거대 자료를 분석한 결과 전력 사용량에 미치는 기상 요소로는 기온과 더불어 습도, 풍속, 체감온도 등인 것으로 나타났다. 또한, 신이레 등 (2016)의 "전력수요예측을 위한 기상정보 활용성평가"에 따르면 기온, 풍속, 습도가 고려된 불쾌지수와 체감온도가 전력수요 예측오차와 상관성이 높았고, 풍속, 습도 등도 기온과 더불어 전력 수요 예측에 있어 고려해야 할 기상변수인 것으로 확인되었다. 불쾌지수 = 1.8 x 기온 - 0.55 x (1 - 상대습도) x (1.8 x 기온 - 26) + 32 로 계산되며, 체감온도 = 13.12 + 0.6215 x 기온 - 11.37 x 풍속**0.16 + 0.3965 x 풍속**0.16 x 기온으로 계산된다. 정현철 등 (2018)에서도 마찬가지로 생활 기상지수인 불쾌지수와 체감온도를 활용했을 때, 각각 하계와 동계의 최대 전력 수요 예측의 정확성을 향상시키는데 기여한 것으로 나타났다.

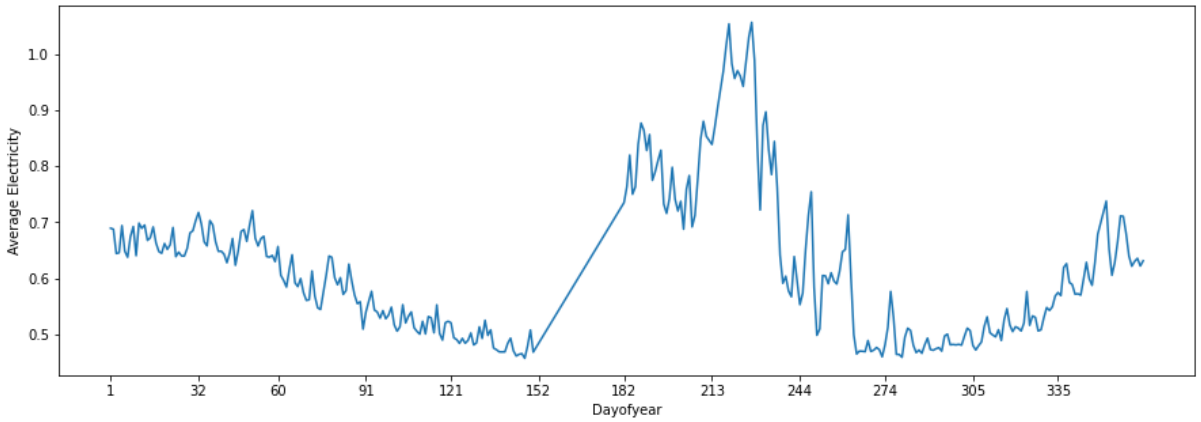
둘째, 날짜 및 시간 변수는 년, 월, 일, 요일, 시간으로 구분하였고, 평일과 주말을 구분하여 주말을 나타내는 변수, 평일 공휴일 및 대체공휴일일 여부를 나타내는 변수를 추가하였다. 평일과 주말은 다른 전력사용 패턴을 보이기에 주말여부를 구분할 필요가 있으며, 명절 연휴와 같은 공휴일과 대체공휴일 역시 상이한 전력사용량을 나타내기에 효과적인 분석을 위해 공휴일 여부를 구분하였다.

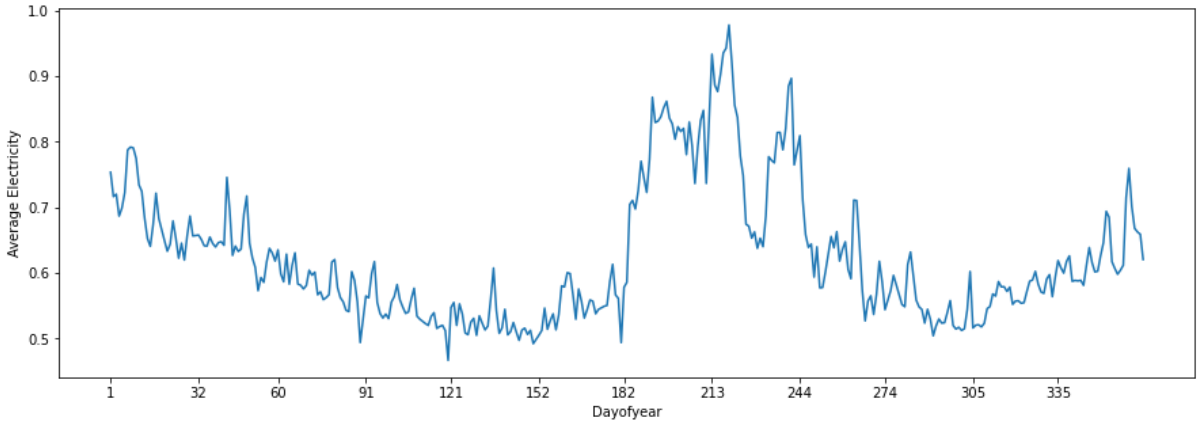
셋째, 코로나 19 관련 변수로는 사회적 거리두기 단계를 활용하였다. 2020 년 1, 2 월 전 세계적으로 코로나 19 바이러스가 퍼지고 코로나 19 가 장기화됨에 따라 사회적으로 많은 변화가 나타났다. 본 분석은 21.1~22.12 의 데이터를 통해 23.1~23.3 의 전력사용량을 예측하는 것이기에 코로나 19 의 영향을 고려할 필요가 있다. 선행 연구들에서 확진자 수를 많이 활용했는데, 코로나 19 가 장기화되고 완화되면서 확진자 수를 직접 사용하는 것보다 확진자 수를 기준으로 시행한 사회적 거리두기 단계를 코로나 19 변수로 활용하는 것이 더욱 효율적이라고 판단하였다. 사회적 거리두기는 코로나 19 에 대한 대응으로 지역사회 감염 확산을 막기 위해 2020 년 3 월부터 시행된 것으로, 사회적 거리두기 단계에 따라 다중이용시설 시간 제한, 인원 제한, 시설 운영 중단 등으로 사람들의 외출 및 생활에 영향을 미치기 때문에 전력 소비 패턴 역시 변하게 된다. 박주환 등 (2021)의 "코로나 19 영향을 고려한 머신러닝 기반의 전력 사용량 예측 알고리즘 구현"에서도 코로나 19 확진자 수와 같은 코로나 19 의 영향을 고려했을 때, 전력 사용량 예측 성능이 높아진다고 확인되었다.

마지막으로, 전력거래소에서 제공하는 공공데이터인 '하루전 발전계획용 수요예측량' 변수를 추가하였다. 해당 변수는 다음 날 필요한 전력 발전량을 계획하는데 사용되는 정보로서, 전력 수요를 예측함에 있어 과거 전력 사용량 데이터, 경제 상황 등의 다양한 요인을 고려한다. 따라서 앞서 주된 예측변수로 포함하였던 기상 관련 변수들과는 다른 성격의 예측자 또한 추가하는 것이다. 이를테면, 상기 수요예측량은 고용 상황, 소비자 동향 등의 경제 지표를 감안하여 계산되므로, 자연스레 가정 전력 수요에 영향을 주는 경제적 상황을 반영할 수 있는 보다 정교한 분석이 될 것이라 판단하였다.

3-3) 탐색적 데이터 분석(EDA)

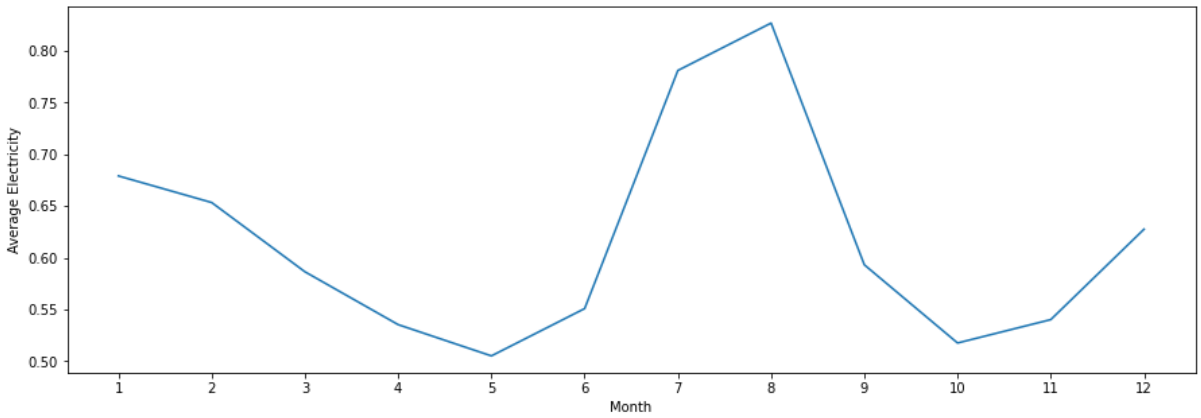
[2021 년, 2022 년 전력사용량]





각각 2021 년, 2022 년의 일별 전력사용량 평균을 나타낸 그래프이다. x 축에 표시된 값들은 매달 1 일에 해당하는 dayofyear 값으로 1=1 월 1 일, 32=2 월 1 일, …., 335=12 월 1 일에 해당하는 값이다. 2022 년의 경우 6 월 데이터가 누락된 점을 감안하고 두 그래프를 비교했을 때, 전반적으로 유사한 패턴을 보이는 것을 알 수 있다.

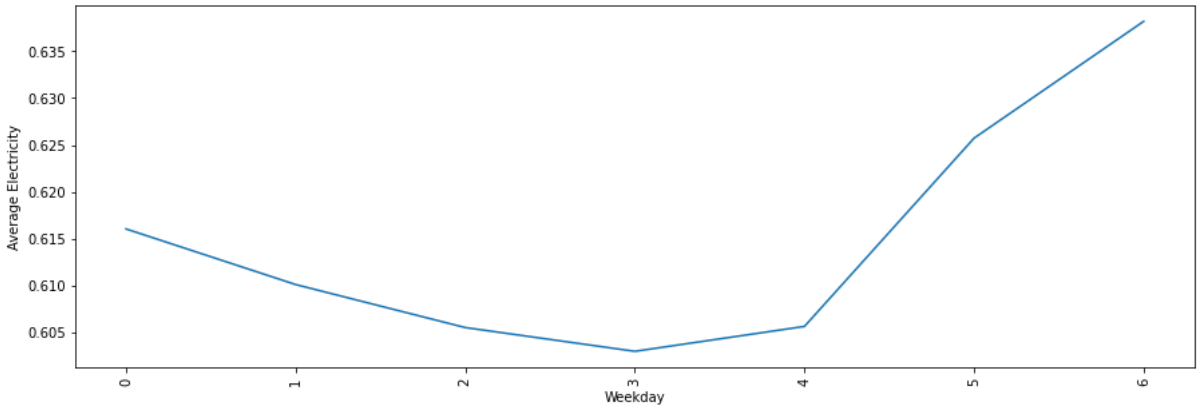
[월별 전력사용량]



월별 2021년, 2022년 전체 전력사용량 평균을 나타낸 그래프이다. 대체로 7~8월 여름에 전력사용량이 가장 높았다가 가을에 감소하며, 12~2월 겨울에도 전력사용량이 증가함을 알 수 있다. 이는 여름에는 냉방 기구, 겨울에는 난방 기구 사용량이 증가하기 때문으로 보인

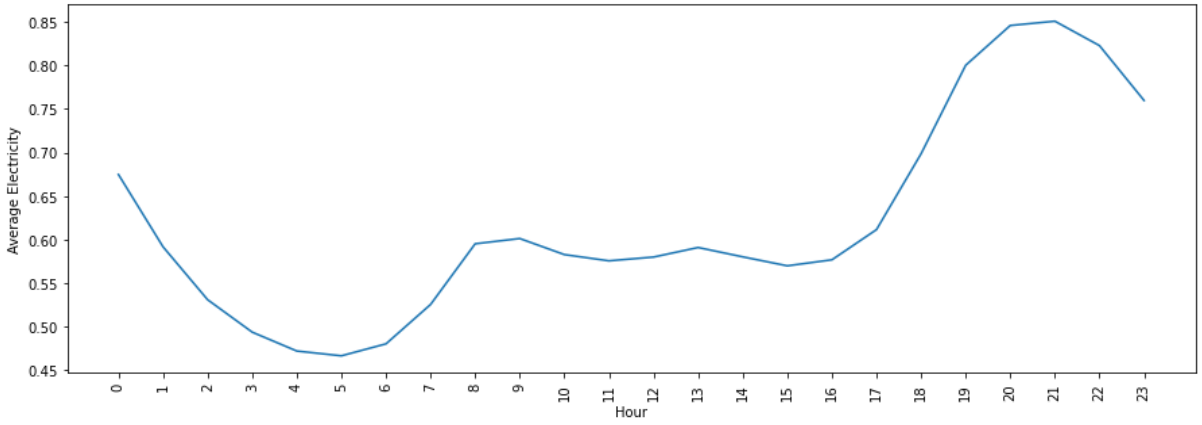
다.

[요일별 전력사용량]



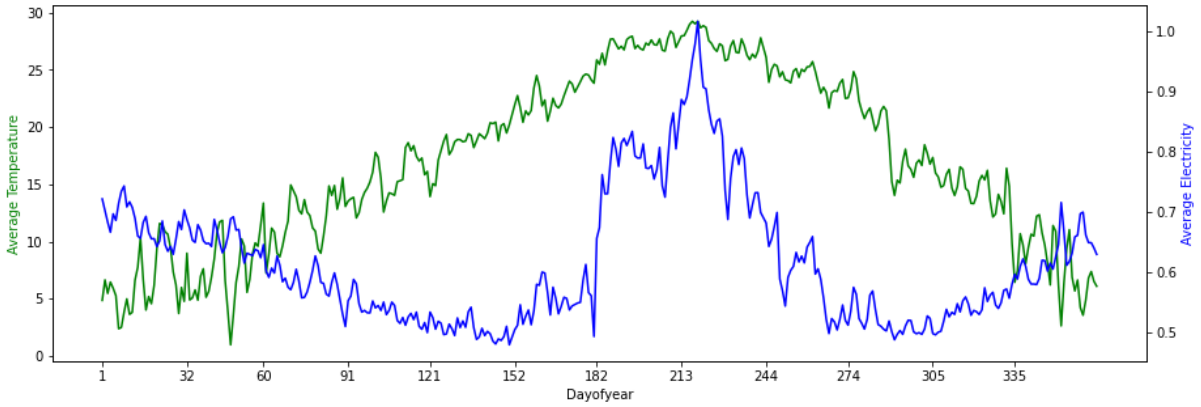
요일별 2021년, 2022년 전체 전력사용량의 평균을 나타낸 그래프이다. x축 weekday는 요일로 0=월요일, 1=화요일, … , 5=토요일, 6=일요일에 해당한다. 제공된 데이터는 주택용 전력으로 보통 평일에는 직장, 학교 등의 이유로 집에 거주하는 시간이 상대적으로 적기 때문에 주말에 전력사용량이 증가하는 것으로 보인다.

[시간별 전력사용량]



시간별 2021년, 2022년 전체 전력사용량의 평균을 나타낸 그래프이다. 마찬가지로 저녁시간인 17~23시에 높은 전력사용량을 보이며, 취침시간인 새벽에 가장 낮은 값을 나타낸다. 0~23으로 표현된 시간의 경우 주기적인 패턴을 강하게 보이기 때문에 사인함수, 코사인함수를 통해 변환한 sin_hour, cos_hour을 시간 변수로 활용하였다.

[일별 기온, 전력사용량]



초록색은 기온의 평균, 파란색은 전력사용량의 평균을 나타낸 그래프이다. 기온의 경우 여름으로 갈수록 높아지며, 겨울은 반대로 낮아진다. 체감온도와 불쾌지수 모두 기온을 활용하여 계산된 값으로 기온과 유사한 패턴을 보이며, 이러한 영향으로 여름철은 전력사용량이 급증하고, 겨울철 역시 전력사용량이 증가한다는 것을 알 수 있다.

3-4) 데이터 전처리

- 제공된 데이터에서 2022 년 6 월이 2022 년 7 월 데이터로 중복 제공되어 2022 년 6 월에 해당하는 행들은 모두 제거하고 분석을 진행하였다.
- 전력사용량 결측인 행 제거: target 에 해당하는 전력사용량이 결측인 행이 5282 개 발견되었다. 이는 전체 데이터셋의 약 0.02% 정도에 해당하는 수준이며, 어떠한 이유로 전력 사용량 측정이 이루어지지 못한 것으로 생각할 수 있다. 따라서 해당 데이터는 차후 구현할 모델의 성능에 영향을 미치지 않을 것으로 판단하여 제거하고 분석을 진행하였다.
- train set 의 전력사용량 outlier 제거: 제공된 데이터의 target 에 해당하는 전력사용량은 대다수가 0~1 내외의 값을 갖는 주택(가정) 단위의 전력사용량에 해당한다. 하지만 일부 고객에서 5000 을 초과하는 값을 갖는 행을 발견할 수 있었고, 이를 포함해 비정상적으로 큰 값을 보이는 이상치들은 일반적인 패턴과는 동떨어진 극단적인 값으로, 오류나 전력 사용량의 잘못된 측정에 의해 나타났다고 보아야 할 것이다. 따라서 이러한 값들은 합리적인 선에서 제거하는 것이 일반적인 상황에서의 예측정확도를 향상시킬 것이라 판단하였다. 또한, 전력사용량은 음수가 될 수 없으며, 본 분석 상황의 경우 0 에 근접한 값보다는 0 보다 훨씬 큰 값들에 한해서 이상치 제거를 하는 것이 타당하다. 따라서, 전력사용량의 90 백분위수(=1.15)를 기준으로 상위 10%에 해당하는 데이터(약 200 만개)들을

제외하고 분석을 진행하였다.

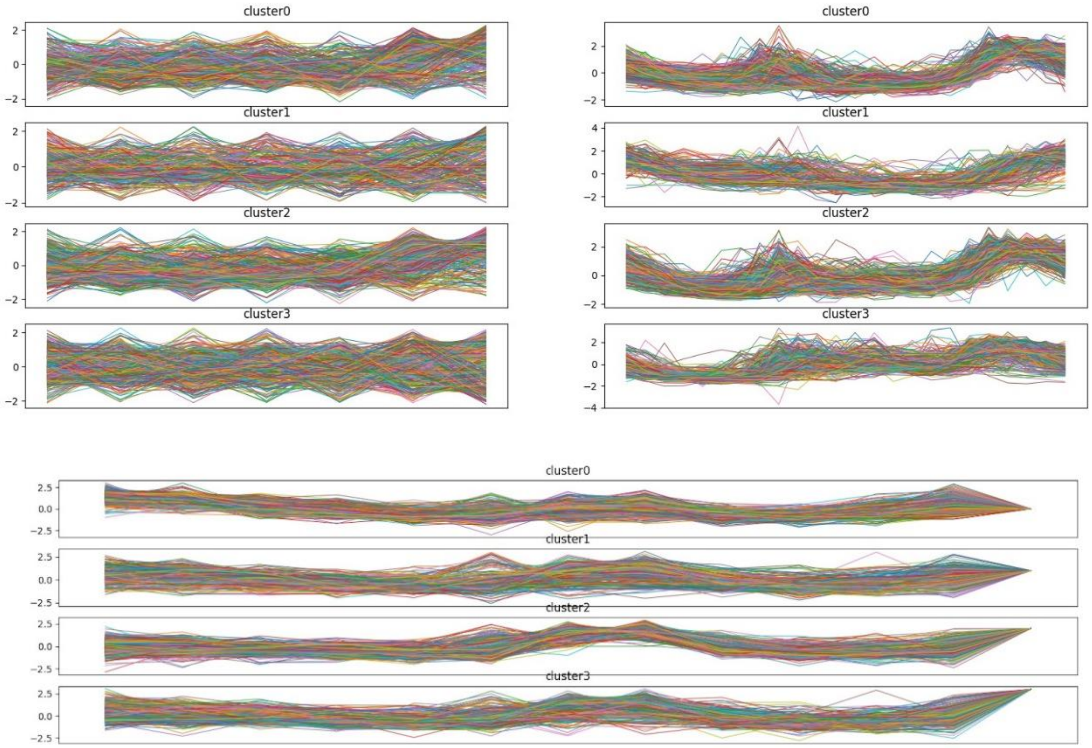
- 사회적 거리두기 변수 one-hot encoding: 앞서 언급한 사회적 거리두기 변수의 경우, 2021~2022 년간 중앙재난안전대책본부(이하 중대본)에서 제주 지역을 대상으로 공식 발표 및 시행한 거리두기 단계와 함께 제주 지자체에서 자체적으로 시행한 거리두기 단계를 모두 고려하여 기간별로 나타내었다. 거리두기 단계의 완화 및 강화 또는 위드코로나 체제로의 전환 이후 거리두기는 각각 규제의 강도가 상이하며, 그에 따라 제주도민의 일상 또한 영향을 받을 것이다. 따라서 one-hot encoding 을 활용해 범주화하여 주어진 데이터셋의 기간 동안 거리두기 단계가 세밀하게 반영되어 예측 모델을 구현하고자 하였다.
- standard_prediction 표준화: 하루전 발전계획용 수요예측량 변수의 경우, 다른 변수들의 scale 과는 다른 4~8 만 범위의 값들을 갖는다. 다른 예측변수들에 비해 절대적으로 큰 값을 가질뿐만 아니라, 그 범위 역시 매우 넓다. 따라서 해당 변수의 과도한 효과를 보정하기 위하여 StandardScaler 를 활용한 표준화를 진행하여, 합리적인 선에서 해당 변수의 상대적 비교가 가능케 하였다.
- client_mean 추가: 고객 1~1500 명의 전력사용량 패턴에 차이가 있기 때문에 전력사용량을 예측함에 있어 고객별 차이를 두기 위해 제공된 target 을 고객 번호별 평균을 구해 client_mean 을 추가함으로써 분석의 성능을 높이고자 하였다.
- 로그 변환 고려: 고객번호별 전력사용량의 편향도를 확인하여 left 또는 right 으로 skewed 된 형태를 보이는 경우 로그변환을 해주려려고 하였으나 전반적으로 고른 형태를 보여 로그 변환을 하지 않았다.

3-5) 분석 알고리즘 및 모델

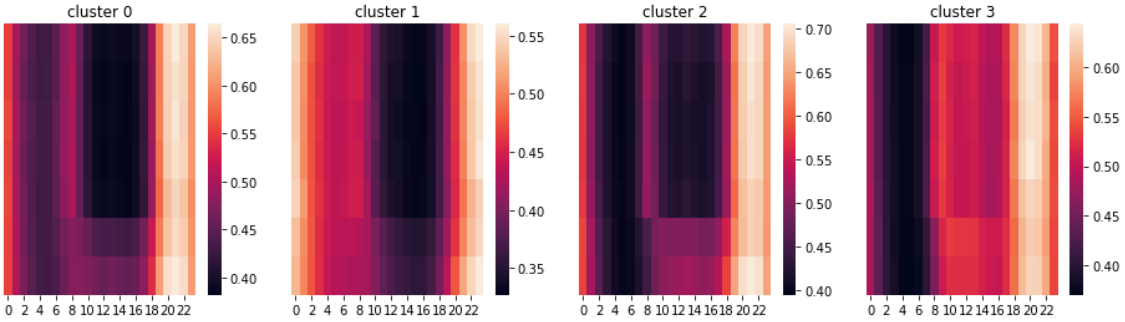
- 클러스터링 기법

  고객의 특성을 예측에 반영하고자 고객의 **요일별, 시간대별, 월별** 전력사용량 추이를 반영하여 Kmeans Clustering 을 수행하였고 이를 변수로 추가해주었다. K-means 클러스터링은 주어진 데이터를 특성 간 거리를 기반으로 K 개의 클러스터로 그룹화하는 알고리즘이다. 군집의 개수는 Elbow method 를 통하여 4 개로 결정하였다. Elbow method 는 데이터에 대해 여러 K 값에 대한 클러스터링을 수행하고, 클러스터 내 오차 제곱합(Sum of Squared Errors, SSE)을 계산하여 K 값의 적절성을 평가하는 기법이다.

또한 군집별 시각화를 통해 각 군집의 특성을 파악하였다. 아래는 군집별로 **요**일별, 시간대별, 월별 전력사용량 추이를 시각화한 결과이다.



아래는 요일에 따른 시간대별 전력사용량 추이 heatmap을 시각화한 결과이다.



시각화를 통해 파악할 수 있는 군집별 특징은 다음과 같다.

군집 0: 저녁 시간대에 전력사용량이 크게 증가한다. 비교적 계절과 무관하게 전력사용량이 높은 편이다.

군집 1: 아침 9 시부터 저녁 9 시까지 전력사용량이 상대적으로 적다가 밤 10 시가 되어서야 증가하는 경향을 보인다. 이를 통해 늦게 까지 일하는 직업(요식업 등)의 종사자로 구성된 가정으로 유추해볼 수 있다.

군집 2: 저녁 시간대에 전력사용량이 크게 증가한다. 여름에 해당하는 6~8 월에 전력사용량이 타군집 대비 두드러진다.

군집 3: 새벽(3~6 시)에는 전력사용량이 크게 줄어들고 타군집에 비해 낮시간에도 전력사용량이다 높은 편이다. 겨울에 해당 하는 11 월~12 월에 전력사용량이 타군집 대비 두드러진다. 아이를 키우는 주부가 있는 가정으로 추측가능하다.

군집별 고객의 수는 군집 2(559 명), 군집 0(379 명), 군집 3(323 명), 군집 1(239 명) 순으로 많다.

- 모델링 과정

  사용한 모델: **GBM(Gradient Boosting Machine)**

  정형데이터 예측에 적합한 머신 러닝 모델을 우선적으로 시도해보았다. 그 중 CatBoost, LightGBM, XGBoost 로 대표되는 그래디언트 부스팅 알고리즘들을 사용했다. 이유는 다음과 같다.

  1. 시계열 데이터: 본 분석에서 사용된 데이터에는 시간에 대한 정보가 중요한 변수로 작용한다. 그래디언트 부스팅 알고리즘들은 이전 시간의 예측 오차를 최소화하도록 모델을 학습하므로 시간적 순서에 따른 패턴을 상대적으로 잘 학습할 수 있으며, 비선형성을 잘 처리하고 복잡한 관계를 모델링할 수 있다.
  2. 트리 기반 모델: CatBoost, LightGBM, XGBoost 모두 random forest 기반의 모델로 범주형 변수 처리에 적합하다. 본 데이터셋에는 dist, weekend, holiday 와 같은 범주형 변수가 포함되어 있기 때문에 트리 기반 모델이 우수한 성능을 보일 것이라고 판단했다.
  3. 대용량 데이터셋: CatBoost 는 대용량 데이터셋에 적합한 자체적인 오버피팅 감소 전략을 가지고 있다다. 이를 통해 모델이 과적합되는 것을 방지하고 일반화 성능을 향상시킨다. LGBM 은 효율적인 리프레벨 학습 방식과 데이터 분할 기법을 사용하여 메모리 사용을 최적화하고 처리 속도를 높인다.
  4. 변수 선택: XGBoost 는 모델을 학습시키는 과정에서 모델의 예측력이 얼마나 감소하는지를 계산하여 변수의 중요도를 파악하는 기능을 제공한다. 이를 활용하여 중요한 변수를 선택하고 모델에 포함시킬 수 있다.

하이퍼 파라미터(초모수) 튜닝: GridSearch 를 사용하여 초모수를 결정하였다. GridSearch 는 가능한 모든 초모수 조합에 대해 Cross-validation 을 통해 성능을 평가하고, 최적의 초모수 조합을 선택한다. 해당 모듈을 활용하여 overfitting 문제를 해결하였다. 초모수 후보군들이 많은 경우 시간이 오래 걸릴 것을 고려하여 본 분석에서는 교차검증 횟수를 5 로 시도해보았다.

LightGBM/XGBoost

1. n_esimators: 생성할 트리의 개수로, 일반적으로 큰 값일수록 모델의 복잡성이 증가한한다. 80, 100, 120 의 값을 시도하였다.
2. max_depth: 트리의 최대 깊이를 제한하는 파라미터로, 과적합을 피하기 위해 사용한다. 3, 5, 7 의 값을 시도하였다.
3. learning_rate: 각 트리의 기여를 조절하는 파라미터로, 학습 속도를 조절하는 역할을 한다. 작은 값일수록 모델 학습이 더욱 느려지지만, 세밀한 조정이 가능해지며 과적합을 완화한다. 0.01, 0.1, 0.2 의 값을 시도하였다.

CatBoost

1. iterations: 부스팅 라운드 수를 의미하며, 100, 200, 300 의 값을 시도하였다.
2. depth, learning_rate: 위와 동일하다.

Train-Test Split: 모델 성능을 확인할 용도로 데이터셋을 8:2 로 train/test data 를 구분하여 학습을 진행하였다.

모델링 시도: 모델링은 크게 모델에 모든 데이터셋을 학습시켜서 예측을 진행하는 방식과 고객번호별 하나의 모델을 구현하는 방식, 두가지 방식을 시도해보았다.

3-6) 분석결과 및 고찰

본 분석에서는 3 가지 머신러닝 기반 모델을 선택하여 예측을 수행하였다. 아래의 표는 각각 단일 모델로 예측을 진행하였을 때의 결과값을 나타낸다.

	LightGBM	XGBoost	CatBoost
RMSE	0.21	0.205	0.198
RRMSE	38.4	37.48	36.2

모델의 성능을 높이고자 각 단일 모델의 결과를 활용하고 그 중에서 가장 성능이 우수한 모델에 한해 grid-search 방법을 활용해 최적의 하이퍼파라미터를 찾아 최종 모델을 구축하고자 하였다.

추가로 1500 명의 고객에 대한 단일 모델링을 시도하였다. 다른 변수들이 동일하게 주어지더라도 고객 각각의 성향만으로도 전력 사용량은 큰 차이를 보일 가능성이 있다. 따라서 고객번호를 중요한 변인으로 보아, 고객에 따라 독립적인 예측을 진행해보는 것이 타당하다고 생각하였다. 마치 random effect model 처럼, 앞서 시도한 모델들 중 가장 예측 성능이 우수했던 CatBoost 단일 모델을 기반으로 1500 명 고객 각각에 대한 모델을 구축하였다. 해당 결과를 요약 및 앞선 결과와 비교하자면, 특정 고객의 전력 사용량이 전체 평균 전력 사용량에 비해 큰 폭으로 크거나 작은 경우, 오히려 앞선 모델보다 뛰어난 성능을 보이는 것을 확인할 수 있었다. 이를 미루어보면, '고객'의 효과가 예측에 있어 유의미한 변화를 가져온다고 볼 수 있다.

그러나 이와 같은 모델은 다분히 subject-specific 하며, 새로운 데이터에 대해 robust 하지 못할 우려가 있다. 또한, 분석 대상 개체가 늘어날수록 구축해야할 모델의 개수 역시 비례하여 증가한다. 따라서, 제주 전역에 퍼져 있는 임의의 고객을 예측하는데 있어서는 그 효율성이 떨어진다고 판단하여 앞서 구축한 모델을 최종 모델로 선정하게 되었다.

한편, 이전에 언급했던 바와 같이, 주어진 전력사용량 데이터의 특성 상 대다수의 target 값이 0~1 사이에 분포하는 것을 확인하였다. 그럼에도, target 값의 분포는 상당한 right-skewed 형태를 띤다. 이는 곧 소수의 outlier 존재만으로도 모델 성능에 큰 영향을 미칠 수 있음을 암시한다. 이를 확인하고자 앞선 전처리 과정에서 수행하였던 것에 더해, target 의 Q3(상위 25%)값을 초과하는 데이터들을 추가로 제거하고 동일한 모델링을 진행해보았다. CatBoost 단일 모형의 경우, RRMSE 값이 28 까지 감소하는 것을 확인할 수 있었다. 다시 말해, target 의 극단값을 일부 제거하는 것만으로도 모델의 성능은 대폭 향상된다고 짐작할 수 있다.

하지만, 예측 모델로서의 안정성과 일관성 및 타당성을 충분히 확보하기 위해서는 모델의

성능에만 집중해서는 결코 안될 것이다. 모델링 과정에서 최종 데이터셋의 크기는 행의 개수가 2000 만을 상회했고 컬럼의 개수 역시 27 로, 비교적 다양한 변수와 함께 충분한 크기의 데이터셋을 활용하여 분석을 진행하였다. 따라서 향후 해당 모델을 활용하여, 보다 일반적인 분석 상황에서 다양한 예측자를 고려한 모델로서 강점이 있을 것이라 판단한다.

3-7) 한계점 및 개선점

본 분석의 한계점은 크게 2 가지로 구분할 수 있다.

- 데이터 안심구역 내 분석 환경의 제한

전력 사용량 데이터는 시간의 흐름에 따라 어떠한 추세를 나타내는 시계열 데이터로, 시계열 데이터의 특성을 고려하여 분석할 필요가 있다. 시계열 예측에 보편적으로 사용되는 ARIMA 와 같은 전통적인 시계열 분석 방법을 고려했으나 해당 데이터의 경우 선형적이지 않고, 평균이 일정하지 않으며, 이상치에 민감한 데이터이기에 전통적인 시계열 방법은 부적절하다고 판단하였다. 따라서, 시계열 예측에 일반적으로 좋은 성능을 보이는 XGBoost, CatBoost, LGBM 과 같은 머신러닝 기반 모델, CNN, LSTM, GRU 등의 딥러닝 모델을 분석에 활용하고자 하였다. 하지만 본 분석과제 1 의 경우 데이터 안심구역 내 주어진 컴퓨터 환경에서만 분석이 가능했는데, 딥러닝 모델에 필요한 tensorflow 및 기타 패키지가 제공된 컴퓨터에 설치되지 않아 딥러닝 모델을 시도하지 못하였다는 한계점이 존재한다. 또한 머신러닝 모델을 학습시킬 때 자동으로 튜닝 및 우수한 성능을 보이는 모델을 제공하는 AutoML 을 활용하고자 하였으나, PyCaret 역시 패키지 설치 오류로 인해 시도하지 못한 한계가 있다.

- 데이터 한계

본 분석에서 주로 활용한 데이터로는 기상 정보, 사회적 거리두기, 하루전 발전계획용 전력 수요예측량이다. 기상 데이터의 경우 정확한 위치 정보가 없어 고산, 서귀포, 성산, 제주 지역의 평균을 구해 활용하였기 때문에 고객별 구체적인 지역의 기상 정보를 반영하지 못한다는 한계가 존재한다. 또한, 코로나 19 의 영향을 나타내는 변수로 사회적 거리두기 단계만 활용함과 더불어 전력 사용량 예측에 있어 경제활동과 같은 보다 다양한 변수를 활용하지 못하였다는 점에서 한계가 있다.

이러한 한계점을 보완하여 개선해야 할 점은 다음과 같다.

먼저, 코로나 19 로 인해 사회적 거리두기 시행으로 인한 인원 제한, 운영 시간 단축 등은 사람들의 생활 패턴에 많은 영향을 미쳤는데, 코로나 19 가 점점 완화되면서 일일 확진자수에 대한 민감도도 줄어드는 경향성이 있는 등 분석 기간 내 코로나 19 의 영향을 반영할 수 있는 다양한 변수를 활용한다.

또한, 전력 사용량에는 기상 정보와 더불어 경제 지표도 영향을 미친다고 일반적으로 알려져 있기에 GDP 와 같은 변수를 추가적으로 활용한다. 또한 고객별 정보가 제한된 데이터로 분석을 했기에 고객들의 특성, 생활패턴, 경제활동 등을 반영하여 분석을 진행한다면 더욱 효과적인 분석이 될 것이다.

마지막으로, 시도하지 못한 LSTM, CNN 등의 딥러닝 모델을 사용할 경우 보다 강력한 성능을 보일 것으로 기대되며, AutoML 을 활용한다면 더 최적화된 튜닝이 가능할 것이다.

IV. 활용데이터

- 제공 데이터
  1~1500 에 해당하는 고객번호에 따른 연(2021~2022),월,일,시간별 전력사용량 데이터

- 공공 데이터 및 외부 데이터

1. 산업통상자원부 산하기관 공공데이터:
    1. 한국전력거래소>주요사업>전력관련정보>하루전 발전계획용 수요예측
       : 전력거래소에서 제공하는 2021~2022 년의 일별, 시간별 발전계획용 전력 수요예측량을 나타낸 데이터이다. 제주 지역의 특성 상, 신재생에너지의 발전이 총 발전량의 큰 비중을 차지하고 있기 때문에, 특정 날짜의 발전계획을 수립함에 있어 그보다 하루 전날 수요량 예측을 실시하게 된다. 이를 통해 효과적인 전력수급이 이루어지게끔 하는데, 이 예측량의 크고작음 및 변화량은 곧 본 분석의 목적인 주택(가정)단위의 전력사용량 예측과도 밀접한 관련이 있을 것이라 판단하여 전체 데이터셋에 추가하여 예측 변수로 활용하고자 하였다.

2. 한국전력공사>전기자료>전력통계>한국전력통계>8.2 행정구역별 발전설비 및 발전량

: 한국전력공사에서 제공하는 한국전력통계(제 92 호)는 발전설비추이와 함께 전국 행정구역별(17 개 시도)발전설비와 발전량을 제공한다. 발전량의 경우, 원자력, 석탄, LNG, 신재생, 유류, 양수, 기타로 구분된다. 본 분석에서 주목한 제주 지역의 신재생에너지 발전 행태를 확인 및 타 행정구역과 비교하고자, 행정구역별로 신재생에너지 발전량을 총 발전량으로 나누어 '신재생에너지 발전 비율'을 구하고, 이를 통해 제주 지역의 신재생에너지에 대한 의존도를 '2.제안배경'에서와 같이 시각화하고자 해당 데이터를 활용하였다.

2. 외부 데이터:

1. 기상자료개방포털>데이터>기상관측>지상>종관기상관측>제주도(고산, 서귀포,성산, 제주) >기온, 풍속, 습도

: 기상자료개방포털에서 제공하는 2021~2022 년의 기온, 풍속, 습도를 나타낸 데이터이다. 분석 데이터의 경우 개인정보 사항이 비식별화 처리되어 제주 지역이라는 정보 외에 제공된 정보가 없어 구체적인 위치를 알 수 없기에 기상자료개방포털에서 제공하는 제주 지역의 고산, 서귀포, 성산, 제주 4 곳에 해당하는 기온, 풍속, 습도 데이터 각각의 평균을 구해 기상 변수로 활용하였다.

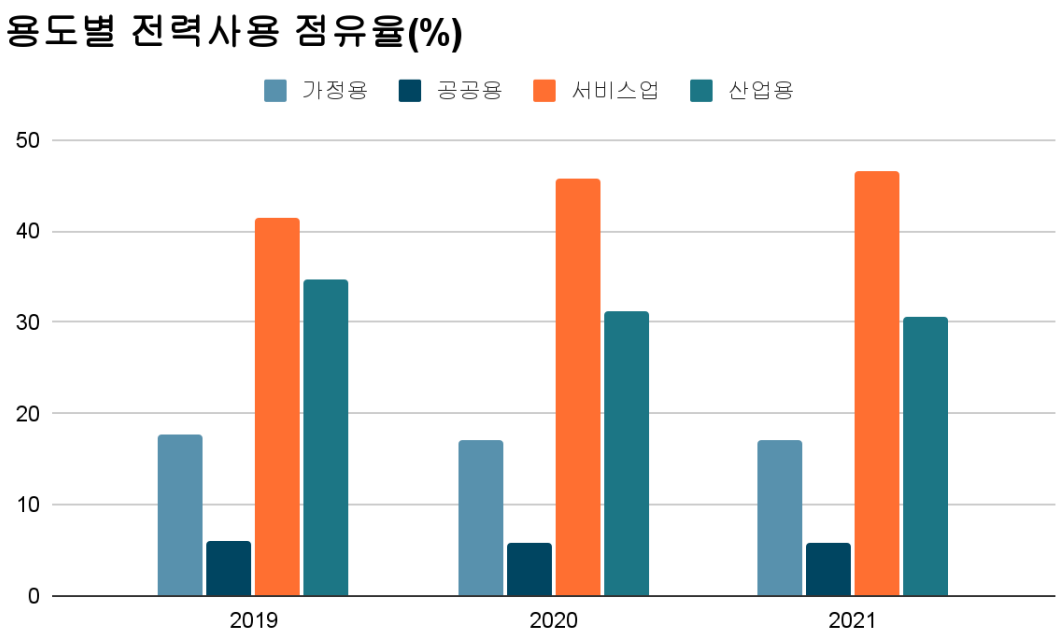
Ⅴ. 사업화방안 및 기대효과

사업화방안:

- '관광 및 서비스업 분야의 전력 수요 예측을 통한 안정적인 관광사업 유지'

본 분석에서 구현한 모델은 기상요소를 중심으로 한 전력 사용량 예측 모델이다. 이를 확장하여, 관광 및 서비스업 분야의 전력 수요량을 파악해 궁극적으로는 제주 지역에 넓게 퍼져 있는 각종 관광명소 및 서비스업에서 보다 원활한 관광사업을 지속할 수 있도록 사업화할 수 있을 것이다. 국가통계포털에서 발표한 지난 3 년 동안(2019~2021)의 '제주 지역 용도별 전력사용량' 통계 자료에 따르면, 4~50%에 달하는 전력을

서비스업에서 사용하고 있을 정도로 서비스업에서의 전력 수요량이 큰 비중을 차지하고 있다.



국내 최대의 관광지인 제주 지역의 특성 상, 많은 관광 및 서비스업 분야에서 냉방 및 난방 부하의 비중이 높게 나타나기 때문이다. 이는 전력수요와 기온 등의 연관성이 제주 지역에서 더욱 두드러지게 나타남을 시사하며, 서비스업이 증가할수록 냉방 및 난방 부하 또한 증가하게 되어, 제주 지역 총 전력 사용량의 높은 비중을 차지하는 관광 및 서비스업에서의 전력 수요 예측은 매우 중요한 문제임을 보여준다. 본 분석에서 제공받은 데이터는 일반 주택(가정)에 한정한 전력 사용량이지만, 이를 발전시켜 관광산업 및 서비스업 분야의 전력 사용량 데이터를 활용하여 예측 모델을 구현한다면 제주 지역의 핵심과도 같은 관광사업을 보다 안정적이고 경제적으로 유지해나갈 수 있을 것이다. 나아가, 상기 산업군에서의 정확한 전력 수요를 예측함으로써, 제주 지역 전체의 전력 수요 역시 보다 정확하게 예측하는데 긍정적인 기여를 할 것이다.

- 환경 친화적인 전기 사용량 조절 앱 개발

매달 말일에 다음달에 사용할 것으로 예측된 전력량과 이를 바탕으로 계산된 예상 전기요금 등을 알려주는 앱을 개발한다. 또한 해당 고객의 과거 전기 사용 추이를 분석하여 전기 사용을 줄일 수 있는 방안을 맞춤화하여 알려준다. 예를 들어, 전년도 여름철에 11 시부터 1 시 사이에 특히나 전기 사용량이 많았던 고객의 경우, 해당 시간대에 에어컨 사용을 줄이라는 알림을 발송한다. 올해 1 월 여론조사기관 조원씨앤아이에서 국민민 2,005 명을 대상으로 에너지 공공요금 인상에 대해 질의한

설문조사에 따르면, 60.4%가 '서민층에 부담이 되므로 최대한 억제해야 한다'는 부정적 반응을 보였다. 이는 많은 사람들이 전기요금 인상에 대해 부담을 느끼는 것으로 보이며, 이 같은 반응을 고려하였을 때 해당 앱에 대한 수요 역시 있을 것으로 생각된다. 이는 자연스레 균형 잡힌 전력 사용으로 이어질 수 있을 뿐만 아니라, 전력 사용량의 감소를 통해 환경적 이점까지 기대할 수 있을 것이다.

기대효과:

1. 육지에 비해 기후가 변화무쌍한 제주 지역의 안정적인 전력 수급 계획에 도움을 줄 것이다. 본 분석에서 사용한 주요 예측변수들은 다양한 측면에서의 기상 요소를 고려하였으며, 주어진 데이터셋의 연도적 특성(코로나 19 사태로 인한 생활패턴의 변화) 역시 포함하였다. 따라서 계절이나 기상요소 등 제주 지역만의 특성뿐만 아니라 예기치 않게 발생하는 사회적 변화를 반영하는 방법 또한 제시한 모델로서, 전력 당국이 보다 정확하게 전력 수요를 파악해 공급량과 더불어 예비 전력을 확충하는데 중요한 참고자료로 활용될 수 있을 것이다.
2. 향후 제주 지역의 신재생에너지 보급과 확대에 있어, 효율적인 설비 가동을 가능케 할 것이다. 신재생에너지의 주요 source 인 태양광, 풍력, 바이오 에너지의 경우 모두 기온을 비롯한 각종 기상 조건이 발전량에 많은 영향을 미친다. 따라서 에너지 발전량의 기복이 심한 편이고, 이로 인해 전력 수급에도 차질을 빚을 우려가 있다. 기상 요소를 중심으로 하여 예측된 전력 수요량을 바탕으로, 신재생에너지 발전을 효과적으로 지속하기 위한 발판을 마련할 수 있을 것이다. 이를 통해 태양광 설비 등의 가동시간을 사전에 조절하고 과부족 없이 전력의 수급을 균형있게 유지할 수 있을 것이라 기대된다.

참고문헌

[1] 김혜민, 김인겸, 박기준, 유승훈, "기온이 전력수요에 미치는 영향 분석", 에너지공학 24.2 (2015): 167-173.

[2] 신이레, 윤상후, "전력수요예측을 위한 기상정보 활용성평가", 한국데이터정보과학회지 27.6 (2016): 1601-1607.

[3] 정현철, 정재성, 강병오, "ARIMA 모델 기반 생활 기상지수를 이용한 동 · 하계 최대 전력 수요 예측 알고리즘 개발", 전기학회논문지 67.10 (2018): 1257-1264.

[4] 박주환, 조영호, 이두희, "코로나 19 영향을 고려한 머신러닝 기반의 전력사용량 예측 알고리즘 구현", 대한전기학회 학술대회 논문집 2021.4 (2021): 211-212.

[5] 정희원, "제주계통의 전력수요예측 및 적정 BESS 용량 산정에 대한 연구", 대진대학교 대학원 석사학위논문, 2017.12

[6] 한국전력공사, "2022 년 한국전력통계", 제 92 호, 2023.5

[7] 산업통상자원부, "신재생에너지생산량", 2021.12

[8] 산업통상자원부, "전력수급동향", 2021.12

[9] 국가통계포털(KOSIS), "용도별 전력사용량", 2019-2021