1. 데이터 EDA 및 전처리 수행 내역

주어진 3가지 데이터를 load하고 cds2_data파일을 분리하는 작업을 거쳤습니다. index가 있는 파일은 예측용으로 사용하고 없는 파일은 train set으로 사용하기에 null값을 -1로 대체하였고 filter를 거쳐 두 데이터를 분리하였습니다.

해당 과정 이후 join 함수를 통하여 주어진 3가지 데이터 파일을 하나로 합쳤습니다.

profile table을 통하여 간단한 EDA 및 결측치 전략을 수립하였습니다.
cunstruct year와 num_floors는 결측치가 너무 많아서 해당 데이터는 이후 train에 사용하지 않기로 결정하였습니다. correlation결과 cunstruct_year변수와 generate 사이에는 강한 상관계수가 있다는 것을 파악하였지만 건물의 건설 연도는 나머지 데이터로 파악할 수 없다고 판단하였습니다.

상관계수를 이용하여 파악한 정보 중 area 변수가 generate와 영향이 크다는 것을 알 수 있었습니다. 건물의 면적이 태양광 설치 면적과 연결되어 있을 것이기에 합당하다는 결론이 내려졌습니다. 또한 온도와 건물의 사용 용도 역시 상관계수가 높다는 것을 파악하였습니다.

이 데이터는 수치형으로 보이지만 범주형인 데이터가 몇 가지 있습니다. 바로 사용 용도입니다. 사용 용도는 범주형 변수로 수치값의 대소관계가 의미가 없습니다. 따라서 해당 값은 one hot encoder 기능을 사용하여 범주형 데이터로 파악할 수 있도록 주의하려고 합니다.

추가 EDA 결과는 다음과 같습니다.
air pressure는 정규분포와 근사한 모양을 나타내며 1015 근처 값을 유지합니다.
area는 대부분의 값이 50000 이하이지만 해당 값 이상의 건물도 확인할 수 있습니다. 이는 전기 생산량과 밀접한 상관관계를 보입니다.

Temperatur는 dew_point와 강한 양의 상관계수를 보입니다.

Rain_hourly는 대부분의 값과 영향이 없습니다. 만약 변수를 컨트롤 해야한다면 해당 값을 제거하는 것을 우선적으로 고려해야 할 것 같습니다.

useage는 0번 건물이 50%를 넘는 관측 값을 나타냈습니다. 0번 건물에서 주로 관찰되었다는 점을 우선적으로 고려해야 될 것 같습니다. 자주 관측된 다른 용도의 건물은 6번과 4번입니다.
user id는 대게 균일합니다. 건물에 따른 발전량 편차는 크지 않을 것으로 생각됩니다.

2. 데이터 모델링 수행 내역

cross validation을 수행하기 위하여 for loop 구문을 사용하였습니다.

regresstion을 위한 모델은 linear regression, random Forest, XGB, Decision Tree, MLP, KNN을 사용하였습니다. 반복 횟수는 9번으로 설정하여 train set을 분리하였고 k fold cross validation을 사용하여 학습을 시도하였습니다. 이 중 MLP는 normalization을 하는 경우 일반적으로 좋은 성능이 나오기 때문에 nomalization을 수행한 결과를 반영하도록 하였습니다.

수행 결과 KNN은 MAE가 34.87, Linear Regression은 369.19, MLP는 162.58, Random Foreset는 14.30, XGB는 34.69라는 결과를 얻었습니다. Random Forest 모델이 가장 좋은 성능을 기록하였지만 XGB를 사용하여 본 예측을 수행하고자 합니다.

이전에 filter로 분리한 값들을 날씨 데이터셋과 빌딩 데이터셋과 함께 결합하였습니다. EDA과정에서 파악한 one hot encoder 함수를 추가하여 usage에 따른 분리를 수행하였고 나머지 데이터들과 함께 모델을 학습하였습니다. 해당 결과를 csv 형태로 변환하여 주어진 값에 대한 결과를 추출할 수 있었습니다.