#1

안녕하십니까. 빵빵이들 조의 팀 프로젝트 중간 발표를 맡은 이정현입니다.

발표 시작하겠습니다.

#2

먼저 저희가 프로젝트에서 사용한 데이터는 총 4가지입니다.

기간은 20, 21, 22년도 1월부터 12월까지로 총 3년치이며 범위는 전국, 단위는 한시간입니다.

기상, 전력 수요량, 태양광 발전량, 풍력 발전량 이렇게 총 4개의 데이터를 조사했습니다.

#3

위 데이터들을 바탕으로 EDA를 진행했습니다.

먼저 전력데이터입니다.

전력 수요량 데이터의 경우, 한여름과 한겨울에, 주말보다는 평일에, 새벽보다는 한낮과 저녁에 더 큰 전력 수요를 보였습니다.

#4

태양광과 풍력 발전량 데이터의 경우, 태양광 발전량은 낮, 밤, 시간에 따른 발전량 차이가 컸지만, 풍력 발전량은 거의 동일했습니다.

계절별 친환경 발전량의 차이는 크지 않으나, 풍력 발전량의 경우 겨울에 높은 값을 기록했습니다.

#5

이후 저희는 어떤 기상요인이 친환경 발전량에 유의미한 상관성이 있을지 조금 더 알아보기 위해

상관관계 분석을 진행했습니다. 분석법으로는 피어슨 분석법을 활용했습니다.

먼저, 기상, 태양광, 풍력 데이터를 한시간 단위로 정리한 기상-친환경 통합 데이터를 구축했습니다.

기상 데이터의 경우 501개, 친환경 데이터에는 16개의 지점이 존재합니다. 이를 친환경 데이터에 존재하는 16개 지점을 기준으로 통합했습니다.

#6

기상 및 친환경 데이터를 통합하면서 생긴 null 값을 제거하여 피어슨 상관성 분석을 진행하였습니다.

풍력 발전량의 경우 풍속 변수와의 상관계수가 0.5로, 유의미한 양의 상관관계를 보여주었습니다.

태양광 발전량의 경우 풍속 변수와 상관계수가 0.27로, 의미있는 상관성을 찾아내지 못하였습니다.

지점별로 변수간의 상관성이 다를 수 있다는 가정 하에 지점별로 나누어 상관분석을 진행한 결과,

대부분의 변수들은 통합 상관분석과 비슷한 양상을 보였으나, 습도의 경우, 태양광 발전량과의 상관계수가 최소 -0.64로, 유의미한 음의 상관관계를 보이는 특이점을 발견했습니다.

#7

다음으로 태양광과 풍력 발전량 예측을 위한 여러 개의 모델 학습을 진행하였습니다. 이전 EDA 과정에서 추려낸 기상요소들과 시간, 위치를 이용한 데이터 셋을 구성하였습니다. 특히 기상 관측값에 대한 데이터는 표준화를 진행, 도시명은 one\_hot-encoding을 진행하여 모델에 입력되기 위한 전처리를 진행하였습니다.

#8

모델의 학습을 위한 train test split으로 8:2를 설정, 아래 4가지 모델로 예측을 진행하였습니다.

Decition tree, multi linear regression, random forest, multi layer perceptron 모델의 예측 결과를 rmse로 출력한 결과입니다.

#9

현재 저희 데이터는 발전이 이루어지지 않는 시간대가 많아 0 값이 많이 분포되어 있습니다. 그렇기에 발전량이 0인 데이터와 0이 아닌 데이터를 랜덤 포레스트를 통해 분류한 뒤, 후자 경우의 결과 지표가 가장 좋은 mlp를 통한 예측을 진행하려고 합니다. 이러한 계단식 모델을 구현, 혹은 여러개의 모델의 결과를 결합하여 더 효과적인 모델을 구축하는 방향도 검토 중입니다.

위와 동일한 과정으로 전력 데이터를 분석한 결과, 모든 요소들을 반영하는 것 보다 시간 변수만 반영한 모델의 결과가 더 정확하였습니다. 따라서 전력 데이터는 시계열 분석을 통한 분석을 진행중에 있습니다.

#10

전력 데이터의 시계열 분석입니다.

원본 데이터의 시계열 분석 가능여부를 판단하기 위해, 시계열 분해를 진행했습니다. 그 결과, 원본 데이터는 추세가 없고, 계절성과 주기성은 있는 것으로 파악되었으며 다음 3가지 조건을 만족하는 것을 알 수 있었습니다.

(PPT 자료) (1. 단기 예측, 2. 계절적 변동 요인, 3. 데이터 수가 50개 이상)에 적합한 ARIMA 모델을 적용하는 방향을 잡았습니다.

ARIMA 모델을 적용의 선행조건이 데이터가 정상성을 만족을 판단했습니다.. 여기서 시계열 데이터의 정상성이란, 시간에 따른 평균과 분산이 일정하며, 시간에 따라 어떠한 경향도 보이지 않는 상태를 말합니다. 하지만 원본 데이터로 ADF 검정, KPSS 검정 결과, 해당 데이터는 비정상 시계열 이라는 사실을 확인했습니다.

정상성을 만족시키기 위한 방법론으로 차분을 이용했습니다. 여기서 차분이란, 연 이은 관측값들의 차이를 계산하여 데이터의 추세와 계절성을 제거하는 방법입니다.

1차 차분 결과, (PPT 자료) 보이는 바와 같이 두 검정에서 모두 정상성을 만족하는 결과를 얻을 수있었습니다.

저희는 최종 발표까지 이 데이터로 학습한 ARIMA 모델을 활용하여 예측을 진행할 예정입니다. 또한, ARIMA 이외의 다른 기법의 적용도 고려하고 있습니다.

5. 앞으의 계획 현재 부족한점

1. 상관도가 낮은 부분을 보완해야함?

풍향

2. 모델 정확도를 높일 수 있는 방법

감사합니다.