대본

1. 안녕하세요 태양복사 예측 프로젝트를 진행한 강예진 박선진 장유진입니다.

2. 발표 순서는 먼저

사용한 데이터 설명드리고,

탐색적 데이터 분석을 한 뒤

모델링하고 결과를 말씀드리겠습니다.

3. 먼저 데이터 설명입니다

4. 저희가 사용한 데이터셋은 NASA 프로젝트에서 활용한 데이터로 케글에서 가져왔습니다

Hawaii 마우나로아 화산위치에서 측정한 태양복사로 2016년 9월 ~ 12월 동안 측정한 data입니다.

로우는 삼만개 정도 됩니다

5. 포함되는 변수는 다음과 같습니다.

kaggle에서 본 변수 분포도 볼 수 있는데요 분포도를 컬럼 전체의 분포도를 통해 알 수 있는 유의미한 정보는 없었습니다

6. 데이터를 보면

unix time, 측정 날짜, 측정 시간, 일출 시간, 일몰 시간

등과 같은 컬럼은 date time 타입으로 되어있는데요

더 세분화 된 분석을 위해 년, 월, 일 / 시, 분, 초로 구간을 나누었습니다

7. 다음은 EDA를 보겠습니다

8. 데이터 간의 상관관계를 쉽게 확인하기 위하여 matplotlib과 seaborn 라이브러리를 활용하여 데이터를 시각화 해보았더니 태양 복사는 온도와 가장 관계가 높은 것으로 나타났고 다른 컬럼들은 큰 상관관계를 찾을 수 없었습니다

대신 추가로, 시간대가 정오에 가까울수록 일사량이 높을 것이라고 유추하였는데,

상관관계 값이 낮게 나온 이유를 명확하게 알고자 시간대 별 일사량을 Scatter plot으로 시각화 해보았습니다.

9. 모두가 예상 할 수 있듯 결과는 그림과 같았습니다.

19시에서 5시사이에는 일사량 값이 0에 가깝고,

정오에 가까울수록 변동 폭이 크다는 사실을 확인하였습니다.

10. 그래서 이를 월 별로 측정하면 어떻게 될지 시각화 해보았습니다.

예상대로라면 겨울에 근접할수록 낮 시간대 일사량의 최저값이 감소해야 하는데,

11월은 그렇지 않았습니다.

저희는 날씨와 관련이 있을 것이라 추측하여 해당 해의 하와이 날씨 자료를 확인해보았습니다.

11. 확인해 본 결과, 11월의 강수량이 가장 낮고, 12월이 가장 높았습니다.

강수량에 따라 일사량의 최저값에 변화가 있다면,

습도와 기압에 따라 일사량이 달라질 수 있지 않을까 생각하였습니다.

12. pyspark의 집계함수로 월 별 기온, 습도, 기압을 측정해보았을 때에도

유독 습도에서 11월과 12월이 차이가 뚜렷하게 나타나는 것을 확인할 수 있습니다.

13. 위의 EDA 결과를 통해 관련 있을 것으로 예상되는 변수로 선택한 feature는 다음과 같습니다

14. 지금까지의 전처리, 시각화를 바탕으로 모델링을 해보았는데요

15. 전체 컬럼으로 Linear regression을 하였을 때, RMSE (평균 제곱근 편차)오차가 194로, 굉장히 높게 나온 것을 볼 수 있습니다. 혹시 다른 모델을 사용하면 RMSE를 줄일 수 있을까해서 Random Forest로 분석해보았습니다

16.

Rmse가 약 50 감소한 것으로 나타났지만

하지만 여전히 오차가 크다고 판단해

위에서 본 시각화에서 일사량의 변동폭 큰 시간대를 골라 모델링을 해보았습니다.

17. 그 결과

9월 181

10월 154

11월 131

12월 123 으로

아쉽게도 이 전과 큰 차이는 없었습니다

그래서 월별로도 한번 나눠봤는데요

18.

9월 138

10월 129

11월 101

12월 93으로

12월로 갈수록 점점 오차 값(rmse)이 감소하고 12월은 세 자리 수에서 두 자리수로 오차가 감소했다는 것을 볼 수 있습니다

더운 달일수록 일사량의 변화 폭이 높기 때문이라는 의견이 있었지만 데이터 분석으로는 판단하기에는 한계가 있었습니다

최종적으로 기존의 rmse 194에서 93까지 약 100정도를 감소시킬 수 있었습니다.

원본 데이터셋에서 일사량의 변동 폭은 1에서 1600 사이이므로,

오차범위를 100 미만으로 감소시킨 것은 충분히 유의미한 결과라 생각됩니다.

19 결과를 분석해보겠습니다

20

먼저 데이터 분석 결과

시간대별, 월별 데이터 분포가 다르게 나타난다는 것을 발견했고,

이를 토대로 태양 복사에 영향을 주는 column 찾아 feature에 추가하여 모델링하였습니다

모델링 결과 Linear regression 보다 Random Forest 사용시 정확도 높아진다는 것을 발견했고,

시간대별로 예측한 값이 전체 데이터로 예측한 결과보다 RMSE 감소한다는 것을 알 수 있었습니다.

21 감사합니다!