산불 진화준비 단계 분석

1. 04\_허건호.csv, 05\_김시형.ipynb 사용

2. 분석에 사용한 feature

- 화재발생\_월, 주말여부, 화재발생시간대, 시군구명, 발화지점, 날씨, 온도, 습도, 발화원인 종합

- 소방서명, 서센터명, 출동소요시간, 현장소방서거리, 현장안전센터거리, ~~현장소방지역대거리~~

3. 현장소방서거리, 현장안전센터거리, 현장소방지역대거리와 출동소요시간 사이의 관계 파악

- 산점도: feature 3개 모두 데이터 포인트들이 상향하는 방향으로 분포 -> 약한 양의 선형관계

- 피어슨 상관계수: 현장소방서거리(0.245447), 현장안전센터거리(0.340478), 현장소방지역대거리(0.167438) -> 약한 양의 선형관계(0에 가까워서 선형 관계가 매우 약하거나 거의 없다고 볼 수 있음)

>> 현장소방지역대거리 0 값이 너무 많아서 삭제

>> DT는 변수들의 값을 기준으로 적절한 분할을 선택해서 데이터를 학습하기 때문에 해당 데이터 수치형으로 유지

4. 소방서명, 서센터명 one-hot encoding

5. DT 코드 동일

6. 비교

산불 발생 단계

RMSE on training set: 1.2202954933216211e-18

RMSE on test set: 36.79251939366661

산불 진화준비 단계

RMSE on training set: 5.7323589160017445e-18

RMSE on test set: 36.950820268547744

- 추가된 feature 중 타겟 변수와 관련성이 부족한 feature가 존재하면 RMSE 증가

산불 발생 단계\_ GridSearchCV

RMSE on training set: 48.837347972787185

RMSE on test set: 68.04064653029715

산불 진화준비 단계\_ GridSearchCV

RMSE on training set: 41.21037627700186

RMSE on test set: 36.89089289668015

- GridSearchCV가 최적의 조합을 탐색하지 못하면 RMSE 증가

7. GridSearchCV 하이퍼파라미터 변경

변경 전

param\_grid = {

'max\_depth': [2, 3, 4, 5, 10, 20],

'min\_samples\_split': [2, 3, 4, 5, 12, 13, 14, 15],

'min\_samples\_leaf': [1, 2, 3, 4, 5],

}

변경 후

param\_grid = {

'max\_depth': [3, 5, 7, 9, 10],

'min\_samples\_split': [2, 4, 6, 8],

}

산불 발생 단계

RMSE on training set: 1.2202954933216211e-18

RMSE on test set: 36.79251939366661

산불 진화준비 단계

RMSE on training set: 5.7323589160017445e-18

RMSE on test set: 36.950820268547744

- 추가된 feature 중 타겟 변수와 관련성이 부족한 feature가 존재하면 RMSE 증가

산불 발생 단계\_ GridSearchCV

RMSE on training set: 0.5165788748262219

RMSE on test set: 36.93909799390914

산불 진화준비 단계\_ GridSearchCV

RMSE on training set: 3.027990893738116

RMSE on test set: 36.914003195867565