저는 좌표를 이용해 거리나 동선 최적화 관련된 문제를 해결하고 싶었습니다.

원래는 좌표가 주어지면 그걸 K-means-clustering처럼 그룹화 후 그룹 내에서 동선 최적화를 하고 싶었으나 계속 찾아보니 이건 저희가 배우지 않은 비지도 학습에 가까워서 급하게 주제랑 데이터를 바꾸게 되었습니다.

그래서 차선책으로 고른 데이터는 샌프란시스코 공유자전거 데이터인데요.
이 데이터터를 가지고, 여러 조건이 주어졌을 때 자전거만의 이동시간을 예측하고자 했습니다.

———————

Statin 데이터입니다.

산점도를 그려보면 이런 형태로 분포되어 있습니다

———

거리 계산하기 앞서 설명드리겠스빈다.

Haversine은 프로젝트 토론에서 코치님이 알려주신 라이브러리인데, 위경도로 거리를 구해줍니다.

거리 구해주고,
근데 station으로 다시 되돌아 온 경우도 있습니다. 이런 경우에는 사실상 거리를 알 수 없기 때문에 이상치로 판단하고 처리했씁니다.

———————

시계열 데이터입니다
몇월이냐에 따라 계절을 나누어주었습니다. 고르게 분포되어 있습니다
연월일을 넣으면 요일을 반환해주는 라이브러리가 있습니다

생각과 다르게 평일의 이용빈도가 주말의 이용빈도보다 높습니다.
통근시간, 낮, 밤으로 시간대도 나눠 새로운 feature를
만들어줬습니다.
출퇴근에 가장 많은 이용을 하고 있습니다.

————

그다음 타겟값인 duration을 알아보겠습니다.

이상치가 너무 큽니다. 그래서 총 duration을 4시간으로
제한해줬습니다.

————

그다음은 날씨입니다.
먼저 zip code가 있는데 이건 지역 우편 코드입니다.
다음과 같은 데이터가 보여집니다.
이상기후 feature를 좀더 명료하게 만져줍니다.
바람 최대 속도와 gust(갑자기 부는 바람, 속풍) 최대 속도는 상관이
있기에 결측값을 채워줍니다.

다른 결측값도 중앙값으로 채워줬습니다.

그 후 최종 trip 데이터가 완성되었습니다

———

이제 이걸로 머신러닝 모델을 만들어보겠씁니다.
기본 모델은 평균으로 구했을 때 다음과 같이 mae ~~~ 나옵니다
머신러닝으로 xgboost regressor을 사용했씁니다

Feature importance를 살펴보면 구독 유무가 가장 큰 영향을
끼치는데 이건 카디널리티가 낮아서 그렇다고 보여집니다. 그
이후로는 출퇴근 시간대, 대각선 거리, 계절 순으로 이동시간에
영향을 미칩니다.

Xgboost 모델의 평가지표는 다음과 같습니다.

기본 모델에 비해 성능이 나아졌으나 r2 결정계수가 1에 한참 못 미쳐 신뢰도가 낮을 것 같습니다.

더 나은 성능을 위해서 방향 feature와 weekday, time feature 를 합하여 만드는 방법이 있을 것 같습니다.
시각화와 PDP 부분을 더 디벨롭하여 완성도를 높여볼 생각입니디ㅏ.