파이썬기반기계학습

# 제주도 도로 교통량 예측 AI 경진대회

2022.10.03 ~ 2022.11.14

DACON

# 목차

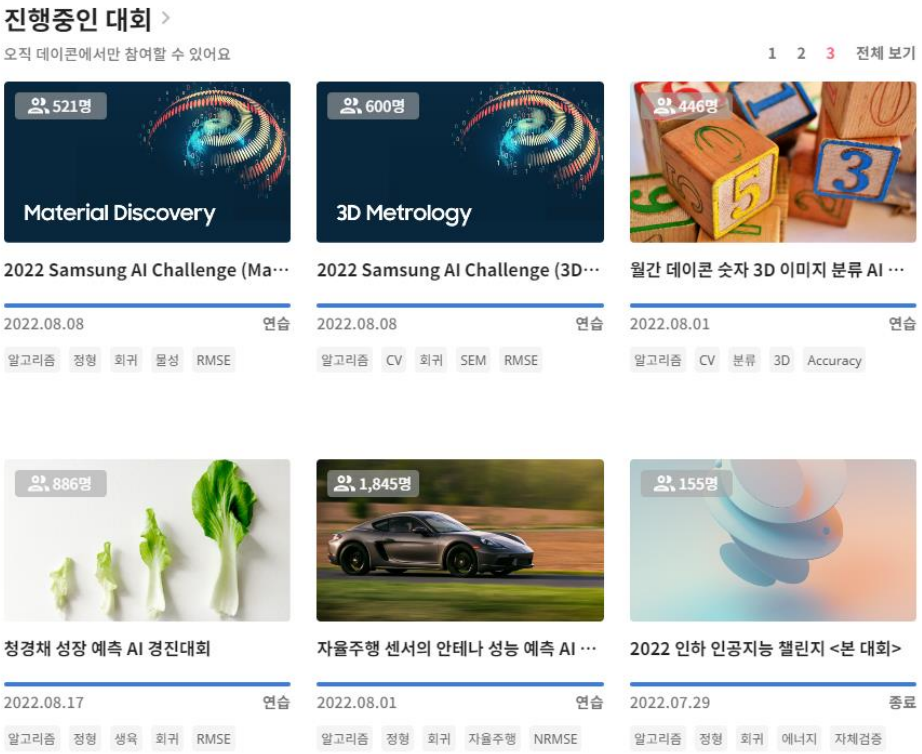
contents

# I INTRO

# I INTRO

[DACON]

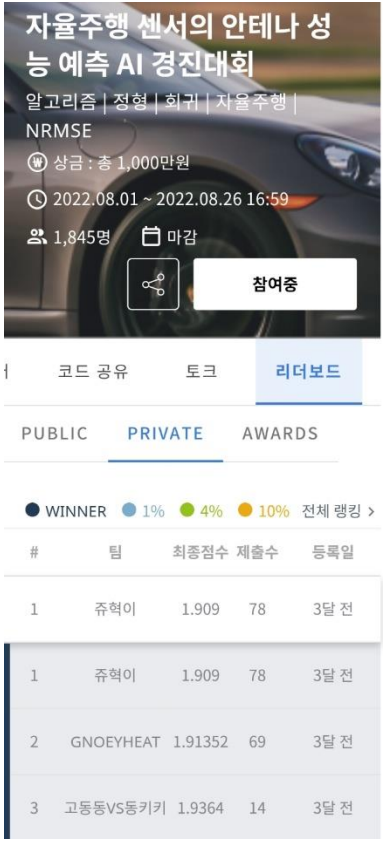
한국판 캐글로 여러 기업들이 주최하며
보통 입상 시 상금과 채용 연계가 이루어짐



# I INTRO

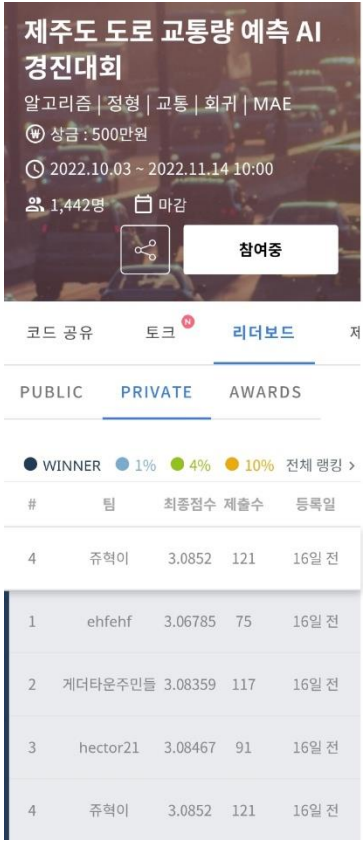
**첫 번째 데이콘 참여**

- LG AI 주최
- 자율주행 센서의 안테나 성능 예측 AI 경진대회
- 1800여명 참가
- Public 1등
- Private 1등
- 최종 x




**두 번째 데이콘 참여**

- 제주도 주최
- 제주도 도로 교통량 예측 AI 경진대회
- 1400여명 참가
- Public 2등
- private 4등
- 최종 3등



# I INTRO

**[배경]**

제주도내 주민등록인구는 2022년 기준 약 68만명으로, 연평균 1.3%정도 매년 증가하고 있습니다.
또한 외국인과 관광객까지 고려하면 전체 상주인구는 90만명을 넘을 것으로 추정되며,
제주도민 증가와 외국인의 증가로 현재 제주도의 교통체증이 심각한 문제로 떠오르고 있습니다.

**[주제]**

제주도 도로 교통량 예측 AI 알고리즘 개발

**[설명]**

제주도의 교통 정보로부터 도로 교통량 회귀 예측

# I INTRO

**[제공 데이터]**

- Train.csv

  2022 년 8월 이전 데이터만 존재하며 날짜, 시간, 교통 및 도로구간 등의 정보와 도로의 차량 평균 속도(target)정보 포함

- Test.csv

  2022년 8월 데이터만 존재하며 날짜, 시간, 교통 및 도로구간 등의 정보 포함

**[외부 데이터]** * 2022년 8월 이전에 수집 가능한 데이터만을 사용

- 국가공휴일.cvs

  2018 ~ 2023 년의 국가 공휴일

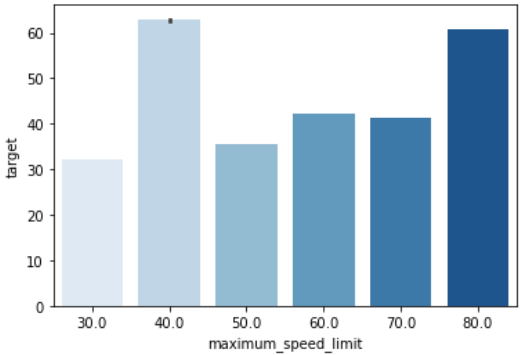
- 제주도장소데이터_20151231.csv (출처 : 공공데이터포털)

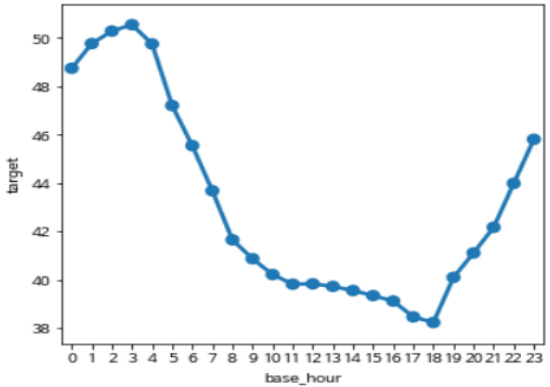
  2015년 제주도 장소데이터로 공항, 항만, 아파트, 마트, 관광지 등의 위치 정보 포함

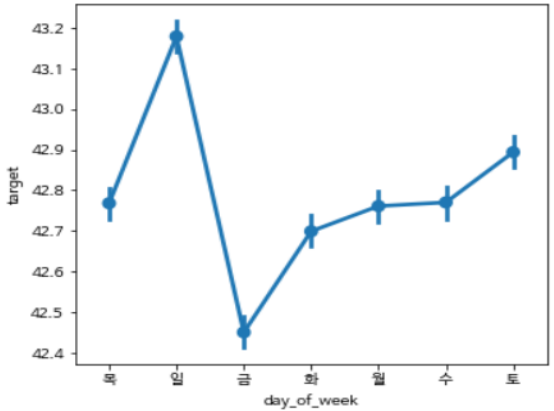
# II EDA

# II EDA

## 1. EDA(일부)



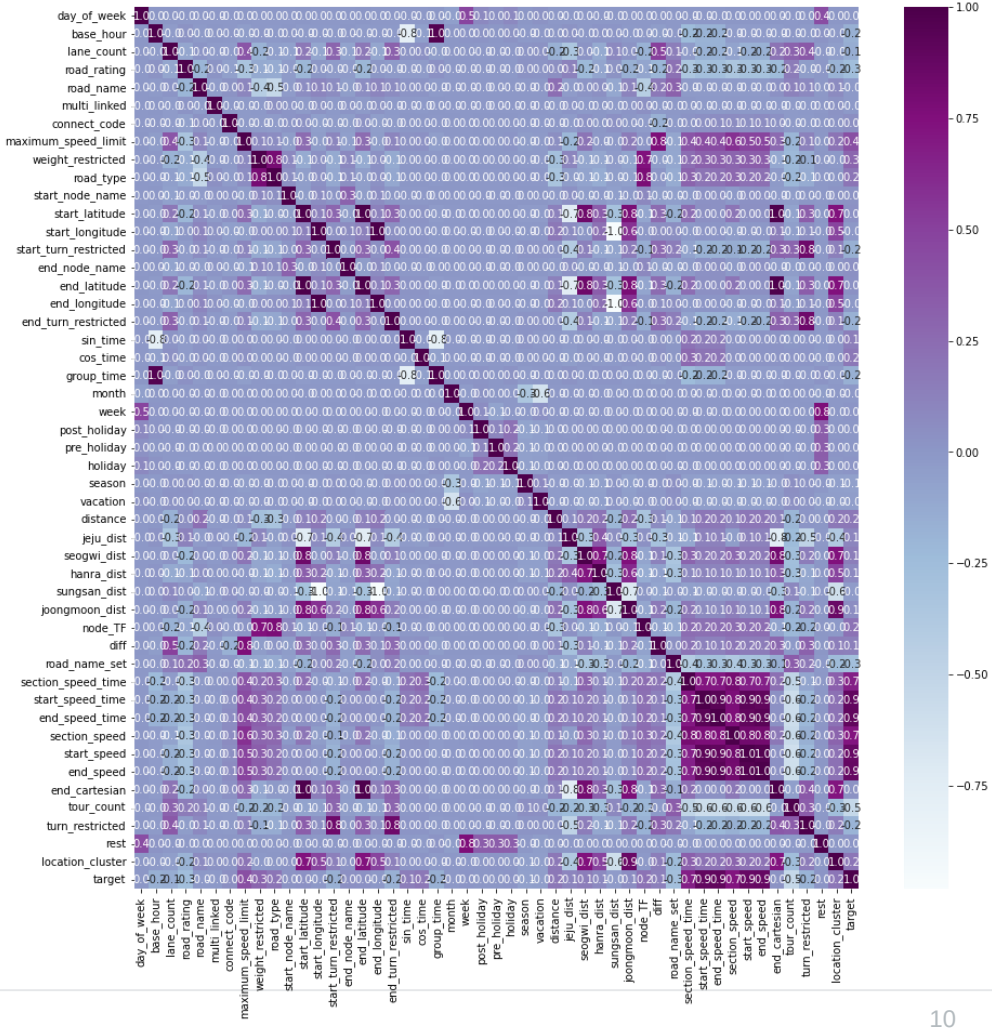

- maximum_speed_limit 값이 40인 경우를 제외하고는 속도제한이 증가함에 따라 target 값도 증가
- 시간에 따른 target값 차이
- 요일에 따른 target값 차이

# II EDA

## 2. 상관관계

**[Target 변수에 대한 상관관계가 가장 높은 변수]**

- Maximum_speed_limit 관련 파생변수
- 관광지 카운트 변수
- 위도경도 관련 파생변수
- 시간 관련 파생변수



# FEATURE ENGINEERING

# III FEATURE ENGINEERING

## 1. DATE 관련 파생변수 생성

| base_date | day_of_week | base_hour |
|---|---|---|
| 20220623 | 목 | 17 |
| 20220728 | 목 | 21 |
| 20211010 | 일 | 7 |
| 20220311 | 금 | 13 |
| 20211005 | 화 | 8 |
| 20210913 | 월 | 7 |
| 20220106 | 목 | 0 |
| 20211213 | 월 | 16 |
| 20211004 | 월 | 15 |
| 20211208 | 수 | 2 |
| 20220623 | 목 | 11 |
| 20220724 | 일 | 2 |
| 20211229 | 수 | 10 |
| 20220507 | 토 | 7 |
| 20220203 | 목 | 16 |
| 20220501 | 일 | 16 |
| 20220701 | 금 | 22 |
| 20211001 | 금 | 17 |
| 20220319 | 토 | 16 |
| 20220210 | 목 | 12 |
| 20220701 | 금 | 21 |
| 20220119 | 수 | 0 |

1. base_date

-> 년/월/일

2. day_of_week

-> 월~일

3. base_hour

-> 0~23시

# III FEATURE ENGINEERING

## 1. DATE 관련 파생변수 생성

| base_date |
|---|
| 20220623 |
| 20220728 |
| 20212020 |
| 20220311 |
| 20211005 |
| ... |

| season | Day_of_week | sin_time | cos_time | group_time | month | week |
|---|---|---|---|---|---|---|
| 3 | 1 | -0.9659 | -2.588e-01 | 2 | 6 | 0 |
| 3 | 1 | -0.70710 | 7.071e-01 | 3 | 7 | 0 |
| 0 | 4 | 0.9659 | -2.588e-01 | 1 | 10 | 1 |
| 2 | 0 | -0.2588 | -9.659e-01 | 2 | 3 | 0 |
| 0 | 6 | 0.8660 | -5.000e-01 | 1 | 10 | 0 |
| ... | ... | ... | ... | ... | ... | ... |

- base_date 에서 Year, month, week, weekdays 파생변수 생성
- 시간과 관련한 피쳐들이 inherently cyclical 하다는 것을 활용하기 위한 sin/cos time 변환
- 새벽, 오전, 오후, 밤 으로 time 그룹핑
- Month 피쳐로 부터 방학시즌(7, 8,12, 1,2월)에 대한 파생변수 생성
- Season 변수 생성

# III FEATURE ENGINEERING

## 2. 좌표 관련 파생변수

| start_latitude | start_longitude | end_latitude | end_longitude |
|---|---|---|---|
| 33.42774727 | 126.662612 | 33.42774877 | 126.662335 |
| 33.50073043 | 126.5291068 | 33.5048113 | 126.5262401 |
| 33.2791451 | 126.3685977 | 33.2800721 | 126.3621475 |
| 33.24608087 | 126.5672043 | 33.2455654 | 126.5662282 |
| 33.46221435 | 126.3265511 | 33.46267677 | 126.3301518 |
| 33.2499487 | 126.5056637 | 33.25218326 | 126.5060688 |
| 33.41841197 | 126.268029 | 33.41417501 | 126.2693776 |
| 33.48239171 | 126.4416217 | 33.48233227 | 126.4422658 |
| 33.25307382 | 126.5063927 | 33.25218326 | 126.5060688 |
| 33.36171667 | 126.7669579 | 33.36433621 | 126.7694089 |
| 33.4194234 | 126.4914948 | 33.42267243 | 126.4929348 |
| 33.24850523 | 126.5697971 | 33.24863305 | 126.5677662 |
| 33.48570693 | 126.6041622 | 33.48005319 | 126.6254858 |
| 33.26411158 | 126.5540433 | 33.26368492 | 126.5509785 |
| 33.31691324 | 126.6246344 | 33.31706538 | 126.6238672 |
| 33.47800071 | 126.5438429 | 33.47744524 | 126.5427027 |
| 33.4858849 | 126.4899786 | 33.48597496 | 126.4864085 |
| 33.45242292 | 126.3826241 | 33.45273432 | 126.3853366 |
| 33.46222797 | 126.4236391 | 33.45796664 | 126.4100768 |
| 33.46531818 | 126.9086122 | 33.47037218 | 126.9027929 |
| 33.50010333 | 126.5128511 | 33.50013221 | 126.512046 |
| 33.27816804 | 126.6676508 | 33.27988259 | 126.6859543 |
| 33.51846796 | 126.6456845 | 33.52766134 | 126.6448212 |
| 33.25194718 | 126.5108937 | 33.25104534 | 126.5105738 |

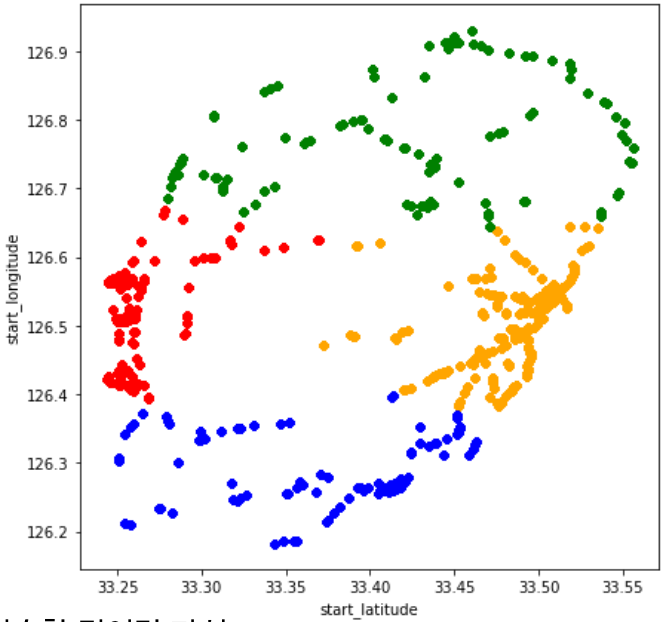
도로의 시작 지점 위도, 경도

도로의 끝 지점 위도, 경도

# III FEATURE ENGINEERING

## 2. 좌표 관련 파생변수

| Location Cluster | distance | jeju_dist | Seogwi_dist |
|---|---|---|---|
| 3 | 0.025711 | 14.555762 | 21.516988 |
| 2 | 0.525891 | 0.737266 | 28.052074 |
| 1 | 0.608399 | 29.022186 | 18.626831 |
| 0 | 0.107352 | 28.436535 | 1.110065 |
| 1 | 0.337949 | 19.092174 | 31.524609 |
| ... | ... | ... | ... |




- 관광지와 가까울수록 교통량이 많을 것/인접한 도로들 사이 교통량이 비슷할 것이란 가설
- 제주시, 한라산, 성산일출봉 등 주요 관광지의 거리 계산 변수 생성
- 출발지와 도착지 사이 거리(harversine) 계산한 변수 생성
- 좌표 기준 4개 구역으로 clustering 한 변수 생성

# III FEATURE ENGINEERING

## 3. 핵심피처 관련 파생변수

| base_hour | road_name | maximum_speed_limit | start_node_name | end_node_name | target |
|---|---|---|---|---|---|
| 17 | 지방도1112호선 | 60 | 제3교래교 | 제3교래교 | 52 |
| 21 | 일반국도11호선 | 60 | 광양사거리 | KAL사거리 | 30 |
| 7 | 일반국도16호선 | 80 | 창고천교 | 상창육교 | 61 |
| 13 | 태평로 | 50 | 남양리조트 | 서현주택 | 20 |
| 8 | 일반국도12호선 | 80 | 애월샷시 | 애월입구 | 38 |
| 7 | 경찰로 | 60 | 시청입구2 | 서호2차현대맨션203동 | 28 |
| 0 | - | 60 | 가동 | 나동 | 39 |
| 16 | 외도천교 | 60 | 외도천교 | 외도천교 | 28 |
| 15 | 경찰로 | 60 | 신성교회 | 서호2차현대맨션203동 | 14 |
| 2 | 일반국도16호선 | 50 | 양수장 | 제2가시교 | 52 |
| 11 | 일반국도99호선 | 60 | 노루생이 | 노루생이삼거리 | 47 |
| 2 | 중정로 | 50 | 선경오피스텔 | 정방수퍼 | 40 |
| 10 | 번영로 | 70 | 명도암교차로 | 버으내교차로 | 60 |
| 7 | 일반국도16호선 | 60 | 서홍교 | 서흥동사무소 | 28 |
| 16 | - | 60 | 송목교 | 송목교 | 58 |
| 16 | 일반국도16호선 | 30 | 아라초등학교앞 | 제2아라교 | 32 |
| 22 | 연동로 | 50 | 그랜드호텔사거 | 흘천5교 | 35 |
| 17 | 중산간서로 | 70 | 중산간서로6091 | 장전1교차로 | 50 |
| 16 | 중산간서로 | 70 | 광령3교차로 | 고성교차로 | 56 |
| 12 | 일반국도12호선 | 80 | 오조한도교입구 | 송내교차로 | 46 |
| 21 | 일반국도12호선 | 70 | 종합운동장입구사거 | 동산교 | 21 |
| 0 | 일반국도12호선 | 80 | 동부장의운수사 | 농업용관정 | 70 |
| 23 | 지방도1118호선 | 60 | 양천동 | 뱅듸왓교차로 | 37 |
| 7 | 새서귀로 | 60 | 한솔고기국수 | 삼주연립101동 | 33 |

타겟값(속도) 이용한
파생변수 생성

# III FEATURE ENGINEERING

## 3. 핵심피처 관련 파생변수

| start_speed | end_speed | section_speed | start_speed_time | End_speed_time | Section_speed_time |
|---|---|---|---|---|---|
| 48.697943 | 50.298219 | 49.105982 | 46.450450 | 48.659670 | 45.269144 |
| 26.400712 | 26.400712 | 47.203323 | 26.562992 | 26.562992 | 35.375781 |
| 59.101720 | 65.118140 | 56.858438 | 60.135135 | 66.588964 | 39.794253 |
| 23.755158 | 25.445418 | 25.030004 | 20.789883 | 23.299611 | 22.146268 |
| 39.873670 | 39.873670 | 51.188650 | 40.518182 | 49.972727 | 41.931997 |
| ... | ... | ... | ... | ... | ... |

- 시작점, 끝점, 도로명에 따른 train data 의 target 변수의 평균값을 이용한 변수 생성

  1 ) maximum speed limit 값에 따른 target 평균 -〉 start_speed, end_speed, section_speed

  2 ) base hour 값에 따른 target 평균 -〉 start_speed_time, end_speed_time, section_speed_time

# III FEATURE ENGINEERING

## 4. 관광지 외부데이터

| ID | X축값 | Y축값 | 구분 | 장소명 | 소재지 | 데이터기준일자 |
|---|---|---|---|---|---|---|
| 3 | 126.5688 | 33.23655 | 교통시설 | 동방파제 | 제주특별자치도 서귀포시 서귀동 758-2 | 2015-12-31 |
| 4 | 126.5626 | 33.23507 | 지명관련 | 새섬 | 제주특별자치도 서귀포시 서귀동 산 3-3 | 2015-12-31 |
| 5 | 126.5997 | 33.23031 | 지명관련 | 섶섬 | 제주특별자치도 서귀포시 보목동 산 1 | 2015-12-31 |
| ... | ... | ... | ... | ... | ... | ... |

제주도장소(POI)데이터_20151231.CSV

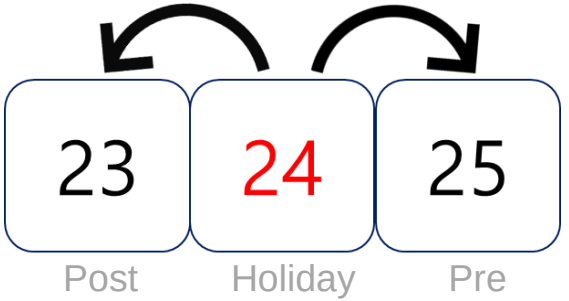
| end_ latitude | end_ longitude | Tour cnt |
|---|---|---|
| 33.427 | 126.662 | 42 |
| 33.504 | 126.526 | 690 |
| 33.280 | 126.362 | 46 |
| 33.245 | 126.566 | 410 |
| 33.462 | 126.330 | 76 |
| ... | ... | ... |

- 관광지 주변에 교통량이 많을 것이라는 가설
- 제주도 장소 외부데이터에서 구분이 관광,문화,레저,공원 인 데이터 추출
- 좌표를 이용하여 도착지 반경 2km 이내 지점인 경우를 카운트 하여 tour_cnt 변수 생성

# FEATURE ENGINEERING

## 5. 공휴일 외부데이터

| Post_holiday | Pre_holiday | holiday |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 1 | 1 | 0 |
| 0 | 0 | 0 |
| 0 | 1 | 0 |
| … | … | … |




- 공휴일에는 교통혼잡도가 달라질 것이라는 가설
- 대체공휴일을 고려하여 공휴일 1일 전, 1일 후 또한 포함하여 파생변수 생성

# III FEATURE ENGINEERING

## 5. 최종데이터

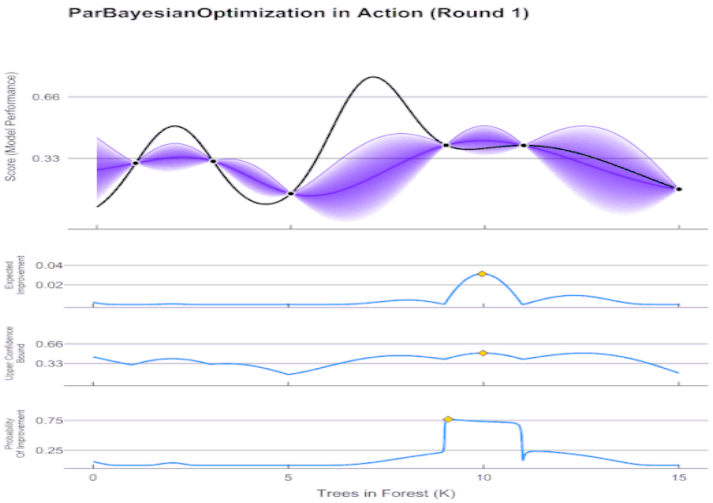
| Day_of_week | sin_time | cos_time | … | start_speed_time | End_speed_time | Section_speed_time |
|---|---|---|---|---|---|---|
| 1 | -0.9659 | -2.588e-01 | … | 46.450450 | 48.659670 | 45.269144 |
| 1 | -0.70710 | 7.071e-01 | … | 26.562992 | 26.562992 | 35.375781 |
| 4 | 0.9659 | -2.588e-01 | … | 60.135135 | 66.588964 | 39.794253 |
| 0 | -0.2588 | -9.659e-01 | … | 20.789883 | 23.299611 | 22.146268 |
| 6 | 0.8660 | -5.000e-01 | … | 40.518182 | 49.972727 | 41.931997 |
| … | … | … | … | … | … | … |

- Feature Engineering 을 통해 총 48 개의 피쳐 생성

# IV MODELING

# IV MODELING

## 1. Basyesian Optimization (Optuna)




OPTUNA

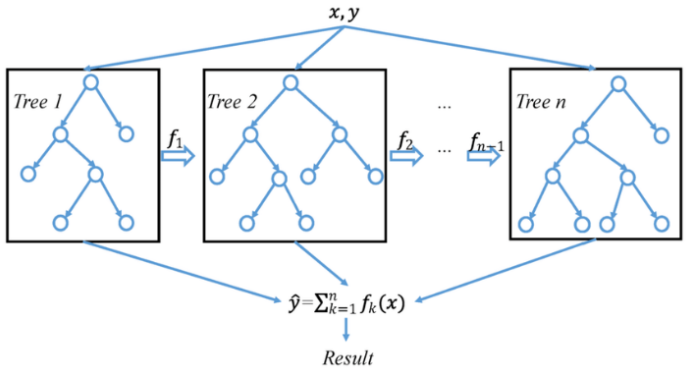
```python
def objective_xgb(trial: Trial, x, y):
    params = {
        "n_estimators": trial.suggest_int('n_estimators', 500, 5000),
        'max_depth': trial.suggest_int('max_depth', 8, 16),
        'min_child_weight': trial.suggest_int('min_child_weight', 1, 300),
        'gamma': trial.suggest_int('gamma', 1, 3),
        'learning_rate': trial.suggest_categorical('learning_rate', [0.008,0.01,0.012,0.014
        "colsample_bytree": trial.suggest_float("colsample_bytree", 0.5, 1.0),
        'lambda': trial.suggest_loguniform('lambda', 1e-3, 10.0),
        'alpha': trial.suggest_loguniform('alpha', 1e-3, 10.0),
        'subsample': trial.suggest_categorical('subsample', [0.6, 0.7, 0.8, 1.0]),
        'random_state': 42
    }
```

- 베이지안 최적화 라이브러리인 Optuna를 이용, 모델 별 최적의 하이퍼파라미터 획득

# IV MODELING

## 2. 모델 실험 리스트

| Models | CV MAE |
|---|---|
| Model Stacking (lgbm+xgb+hist+cat) | 2.77 |
| XGBoost | 2.94 |
| CatBoost | 2.95 |
| histGradientboost Regressor | 3.03 |
| Gradientboost Regressor | 3.11 |
| LightGBM | 3.26 |

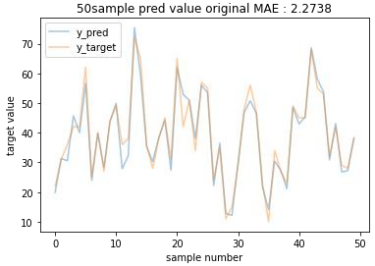



**"최종모델 XGBoost 선정"**

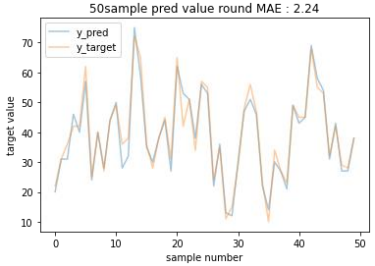
- 각종 모델 테스트 결과 및 추론속도를 고려, 최종모델 XGBoost 선정
- 단일모델을 사용함으로써 실제 산업에 적용시 빠른 평균속력 추론이 가능

# IV MODELING

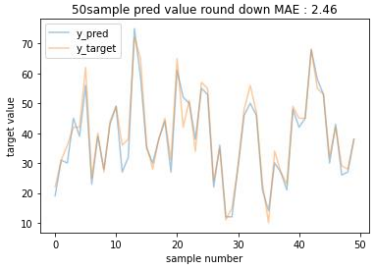
## 3. Post Processing ( 정수형 변환)



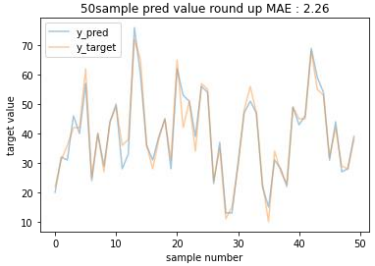

원본 CV MAE




반올림 CV MAE




내림 CV MAE




올림 CV MAE

| 형변환 | CV score |
|---|---|
| 원본 | 2.9848 |
| 반올림 | 2.9731 |
| 올림 | 2.9795 |
| 내림 | 3.0369 |

"round값 변환선정"

- trainset의 타겟값이 정수형이므로 형변환시 MAE값 측정 테스트
- 샘플 뿐 아니라 전체데이터셋 Fold별 CV결과를 반올림, MAE 측정 시 성능향상을 확인

# VI OUTRO

# V OUTRO

## 1. Propose

**[지도 API 활용]**

- 각 도로 별 일정 범위 내의 관광지 수 피처 생성
- **Data-Leakage**로 인해 사용하지 않았으나 성능향상 확인
- 실제 미래 예측 모델에서는 사용이 가능할 것

**[날씨 예보 데이터]**

- 날씨에 따른 Target값 영향이 존재함
- Data-Leakage로 인해 날씨 데이터는 사용 X
- 실제 미래 예측 모델에서는 사용이 가능할 것

**[Stacking]**

- 예측 성능 향상 가능

# Thank you

DACON