https://dacon.io/competitions/official/235985/codeshare/7038?page=1&dtype=recent

주제

2022년 8월 이전의 날짜, 시간, 교통 및 도로구간 등의 정보를 가지고 2022년 8월의 차량 평균 속도(km) 예측

데이터

train.csv

- 2022년 8월 이전 데이터만 존재 (단, 날짜가 모두 연속적이지 않음)
- 4,701,217개의 데이터
- id : 샘플 별 고유 id
- 날짜, 시간, 교통 및 도로구간 등 정보
- target : 도로의 차량 평균 속도(km)

test.csv

- 2022년 8월 데이터만 존재 (단, 날짜가 모두 연속적이지 않음)
- 291,241개의 데이터
- id : 샘플 별 고유 id
- 날짜, 시간, 교통 및 도로구간 등 정보

EDA

Q1. 위도/경도별로 분포는 어떠한가? → 위경도 좌표 plot

Q2. 차선수에 따른 속도 차이는?→ 차선수에 따라 차이가 큼

Q3. 주요 명소/공항은 교통이 집중되지 않을까?→ 17, 18시 낮은편

Q4. 주중/주말 평균 속도의 차이는? → 차이가 없음

Feature Engineering

- **Drop original feature**

  'vehicle_restricted', 'id', 'height_restricted'

- **Create new feature**

  - 제주 공항까지의 거리(km)
  - 한라산까지의 거리(km)
  - start_node_name과 end_node_name을 key값으로 만들어 LabelEncoding
  - 위경도 좌표만으로 Clustering(KMeans) : Clustering Plotting 결과 군집 수가 6일 때 각 좌표가 명확히 구분되어 6으로 설정
  - 공휴일 전후 1 ~ 2일 여부 : 일반적인 공휴일 기준으로 전후 1 ~ 2일을 기간을 더 두어 binary화

- 최고 제한 속도로 도로 주행시 소요 시간
- 각 도로의 start, end node의 위경도 좌표로 해당 도로의 방위각 계산
- 일반적인 요일 순서대로가 아닌 LabelEncoding으로 진행
- 시작 노드 == 종료 노드 여부

Modeling

- lane_count를 1, 2, 3으로 나누어 모델링
- LGBM, XGBoost는 optuna로 파라미터 튜닝
- Ensemble - LGBM : XGBoost : CatBoost = 0.65 : 0.25 : 0.1

차별점 및 새롭게 얻은 인사이트

- 차선수에 따라 차이가 크다는 사실을 먼저 파악하여 아예 차선수를 구분한 채로 각각 모델링한 후 합친 점이 인상적이다. (차선수에 따라 하이퍼 파라미터도 각각 최적화함.)
- 새로운 변수를 생성할 시 과적합 우려를 주의해야하겠다.