DACON 수상작리뷰
주차수요 예측 AI 경진대회 - DACON by 한국토지주택공사

[주차수요 예측 AI 경진대회]

◈ 주제 및 문제점

👉 유형별 임대주택 설계 시 단지 내 적정 주차 수요 예측

- 주차 대수 ← 법정주차대수 vs. 장래주차수요
- 장래주차수요: 주차원단위, 건축연면적 기반으로 산출
- 주차원단위 측정 시, 오차발생 & 조사 시점과 실체 건축시점의 시간차 문제(과대, 과소 산정)

◈ 데이터

- train.csv : 단지코드, 총세대수, 임대건물구분, 지역, 공급유형, 전용면적, 전용면적별세대수, 공가수, 신분, 임대료보증금, 임대료, 도보10분거리내지하철역수, 도보10분거리내버스정류장수, 단지내주차면수, **등록차량수 (target)**
- test.csv : 단지코드, 총세대수, 임대건물구분, 지역, 공급유형, 전용면적, 전용면적별세대수, 공가수, 신분, 임대료보증금, 임대료, 도보10분거리내지하철역수, 도보10분거리내버스정류장수, 단지내주차면수
- age_gender_info.csv : 지역 임대주택 나이별, 성별 인구 분포 (10대미만~100대)
- sample_submission.csv : 단지코드, **등록차량수 (target)**

◈ 코드 리뷰

(1) 데이터 탐색, 전처리

- 🔷 Data Load: age_gender, train, test
- 🔷 임대보증금, 임대료 칼럼의 결측치 '-' → NaN 변경
- 🔷 임대보증금, 임대료 float 변환

(2) Feature Engineering

- 🔷 버스 정류장, 지하철 피처
  - test 데이터셋에 버스정류장 값이 50인 데이터 9개 존재 → train의 해당 피처 평균값으로 대체
- 🔷 임대료, 임대보증금 피처
  - **Data Leakage 문제**
  - 결측치 → train 평균값으로 대체

💡 **Data Leakage (정보의 누설)**

- **발생하는 경우**

① test 데이터의 노출: 알 수 없는 정보가 예측에 반영되었을 때, 시간 또는 순서의 개념이 도입되어 있는 부분에서 자주 발생
② train과 test 데이터셋을 통합했을 때
③ target 값과 거의 동일한 파생 변수가 학습에 사용되었을 때 ex) 암 환자 여부 - 악성 암 세포 여부

- test 데이터셋의 통계치는 test 데이터에서 반영될 수 없음

🔷 자격유형 피처: 결측치 → 'A', 'C' 로 대체
🔷 공급유형+자격유형 피처 생성 BUT 기대만큼 고유값 많이 X
🔷 미성년자 비율 join (비율 높을수록 등록차량 수 낮을 것이라는 기대)

🔷 데이터 병합 (train_agg, test_agg)

🔷 임대 건물 구분, 전용면적 별 세대수, 임대보증금: 처리가 애매한 값 → 단지별 평균 or 고유값 개수 대체
🔷 범주형 피처 처리 def reshape_cat_features(data, cast_col, value_col): 함수 사용
  - data에서 cast_col 에 있는 범주형 값들을 각각 칼럼으로 변환 → 지정한 기준이 존재하면 value_col 의 값을, 그렇지 않으면 0으로 채움
    ⇒ 지정한 기준(단지코드)별로 범주형 피처들이 열로 펼쳐진 형태 (one-hot pivot table)
🔷 최종 병합 데이터 (train_data, test_data)
🔷 로그 변환

(3) 모델링

🔷 **Pycaret 중 성능 상위 5개로 블렌딩**
  - 과적합 고려 → 모델 튜닝 X, 디폴트값 사용
🔷 성능 지표 확인, 피처중요도 시각화

✅ PyCaret

- 머신러닝을 빠르고 간편하게 구현할 수 있게 하는 오픈소스 라이브러리
- 사이킷런 기반의 로우코드 머신러닝 라이브러리
- 장점: 편리함, 모델생성과 성능비교까지 한 줄 코드로 자동화 등
- 단점: 복잡한 커스터마이징 한계, 에러 핸들링 어려움 등

◈ **데이터 오류 해결방법**

- 전용면적별 세대수 합계 =/= 총세대수 ⇨ 가장 큰 오류를 보이는 단지 4개 삭제
- 동일한 단지에 단지코드가 2개로 부여 ⇨ 오류 데이터 모두 삭제
- 기입 실수로 인한 매칭 오류 발생 ⇨ 오류 데이터 모두 삭제

◈ 다른 수상작

- 2위: 논문을 참고한 많은 파생변수 생성 → 앙상블 (Gboost, Lasso, ENet)
- 3위: 다항회귀, Cook's D 를 활용한 가중치

◈ **배울점**