머신러닝 프로젝트
1차 트러블슈팅
(3조)

김도겸
류승환
임현수

목차

데이터 소개

데이터 소개

Y Label : fraud_YN (사기 여부, (2))

X Label : 총 24개 칼럼 (test_set 열 1개 포함)

유저 관련 칼럼 (6개)

- age_group (5)
- has_previous_accident (2)
- cumulative_use_count (2)
- b2b (2)
- socarpass (2)
- socarsave (2)

이용/사고 관련 칼럼 (17개)

- car_model (5)
- sharing_type (2)
- accident_ratio (연속형)
- pf_type (3)
- start_hour (6)
- duration (5)
- accident_hour (7)
- repair_cost (연속형)
- insure_cost (연속형)
- accident_location (6)
- car_part1 (2)
- car_part2 (2)
- repair_cnt (연속형)
- acc_type1 (5)
- insurance_site_aid_YN (3)
- police_site_aid_YN (3)
- total_prsn_count (7)

※ 각 칼럼명 우측 괄호 안의 숫자는 범주 수를 의미함

데이터 소개

No.	컬럼명	의미	컬럼 내용
1	fraud_YN	사기 여부	- 0: 정상 - 1: 비정상
2	car_model	차종	- 1: 경형, 소형, 소형SUV - 2: 준중형, 준중형SUV, 중형 - 3: 대형, 승합, 준대형, 중형SUV - 4: 수입 - 5: EV, RV
3	sharing_type	이용 유형	- 0: 왕복 (직접 타고 갖다놓는 것 의미로 추정) - 1: 기타 (쏘카, 타다, 부름 등 서비스 의미로 추정)
4	age_group	연령대	- 1: 21세 이상 ~ 23세 미만 (21, 22) - 2: 23세 이상 ~ 27세 미만 (23, 24, 25, 26) - 3: 27세 이상 ~ 31세 미만 (27, 28, 29 ,30) - 4: 31세 이상 ~ 41세 미만 (31 ~ 40) - 5: 41세 이상 (41 ~)
5	has_previous_accident	누적사고유무	- 0: 누적사고 0건 - 1: 누적사고 1건 이상
6	cumulative_use_count	누적대여횟수	- 1: 1회 - 2: 2~5회 - 3: 6~10회 - 4: 11회~
7	b2b	법인이용	- 0: 개인 - 1: 법인 - 2: 법인구성원
8	accident_ratio	과실률	- 0~100의 수치
9	pf_type	보험료 타입	- 1: PF5% - 2: PF30% - 3: PF70%
10	socarpass	쏘카패스	- 0: 쏘카패스 없음 - 1: 쏘카패스 있음
11	socarsave	쏘카세이브	- 0: 쏘카세이브 없음 - 1: 쏘카세이브 있음
12	start_hour	이용시작시간	- 1: (0,1,2,3,4,21,22,23) 등 비추류 시간 - 2: (17,18,19,20) 등 퇴근시간 - 3: (5,6,7) 등 이른 아침 - 4: (8,9,10) 등 출근시간 - 5: (11,12,13) 등 점심시간 - 6: (14,15,16) 등 오후 한가운데

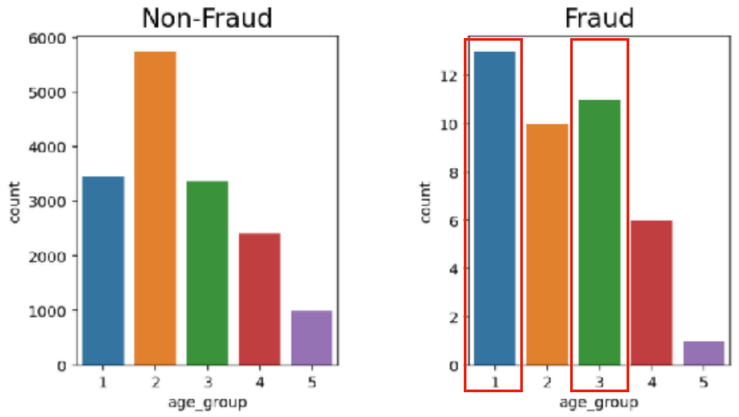
(원본 파일 다운로드 링크 →)

EDA 결과

1) 유저 관련 칼럼 EDA 결과

연령대와 누적사용횟수, 쏘카세이브에서 정상군과 사기군 간의 차이가 나타남

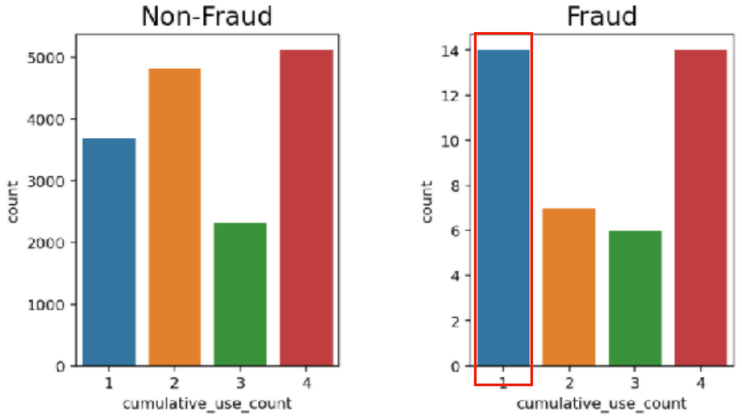
age_group



(1부터 차례대로 ‘21~22세’, ‘23~26세’, ‘27~30세’, ‘31~40’세, 41세~’임)

사기군에서 20대 초반과 후반의 비율이 높게 나타남
(각각 약 10%p, 5%p 높게 나타남)

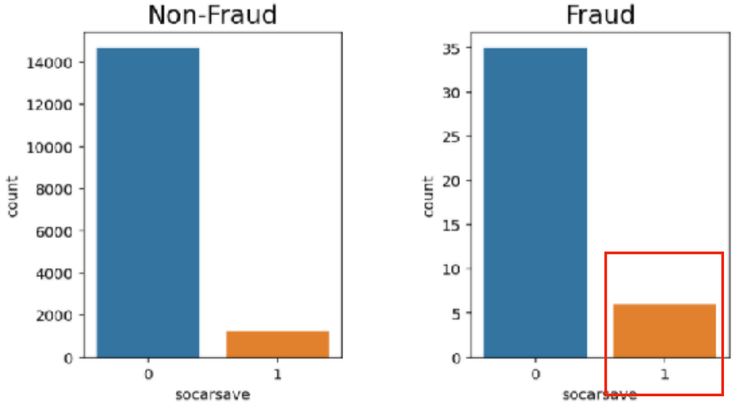
cumulative_use_count



(1부터 차례대로 ‘1회’, ‘2~5회’, ‘6~10회’, ‘11회~’임)

공통적으로 ‘11회 이상’에서 가장 높은 비중을 보이는 한편, 사기군에서 ‘1회’ 비율이 약 10%p 높게 나타남

socarsave



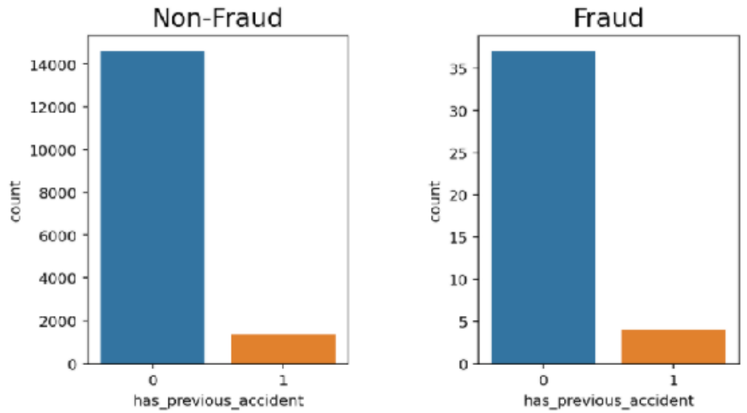
(0이 ‘없음’, 1이 ‘있음’)

사기군에서 ‘쏘카세이브 있음’의 비율이 2배 가까이 높게 나타남
(각각 약 8%, 15%)

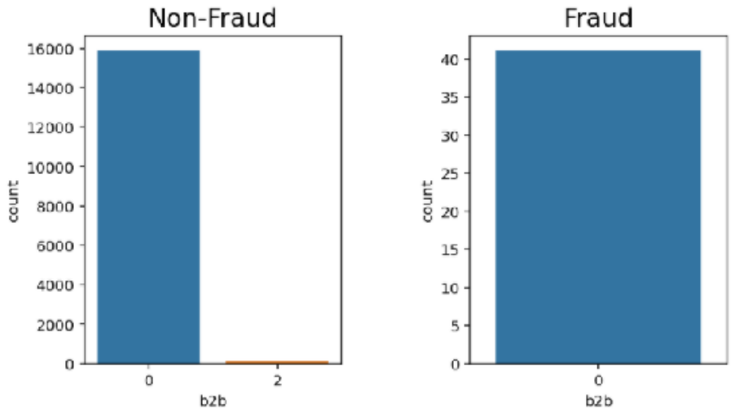
1) 유저 관련 칼럼 EDA 결과

누적사고유무와 b2b, 쏘카패스에서는 두 그룹 간 유사한 양상을 보임

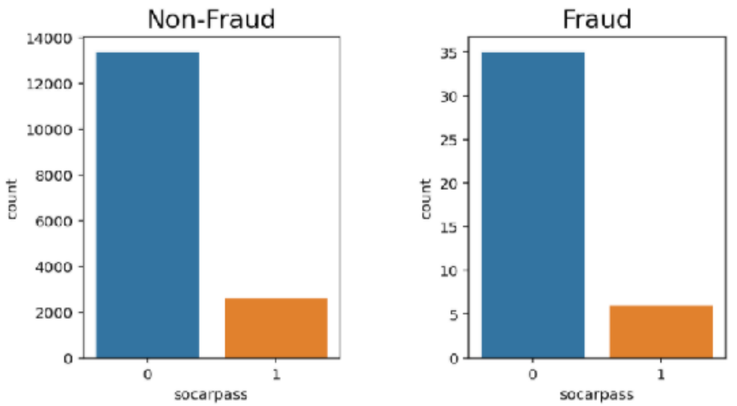
has_previous_accident



b2b



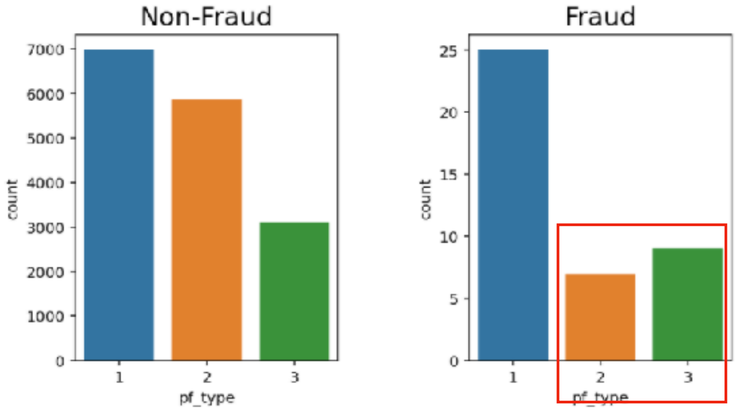
socarpass



2) 이용/사고 관련 칼럼 EDA 결과

6개 항목 (자기부담금, 이용시작시간, 대여기간, 사고시각, 사고위치, 전면손상 여부)에서 두 그룹간 차이를 보임

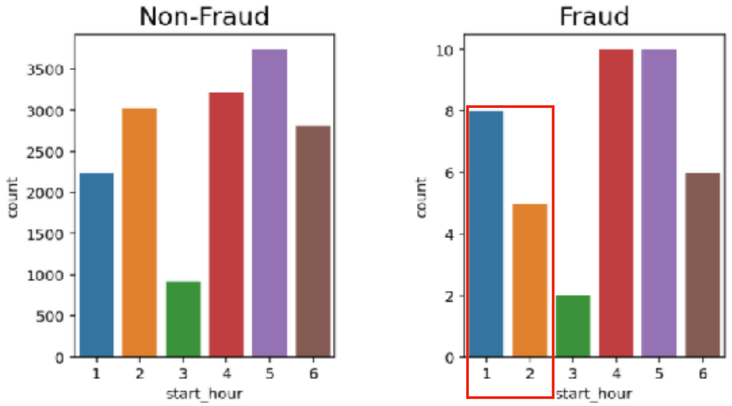
pf_type



(1부터 차례대로 ‘5%’, ’30%’, ’70%’임)

사기군에서 자기부담금이 가장 적은 보험에 가입하는 비율이 높게 나타남
(각각 1의 비율이 44%, 61%)

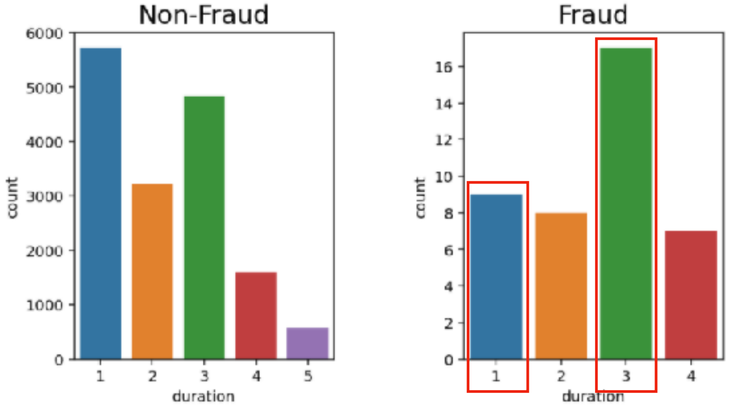
start_hour



(1부터 차례대로 ‘21~04시’, ‘17~20시’, ‘5~7시’, ‘8~10시’, ‘11~13시’, ‘14~16시’임)

사기군에서는 비교적 심야시간 대여 비율은 높고, 퇴근시간 대여 비율은 낮게 나타남

duration



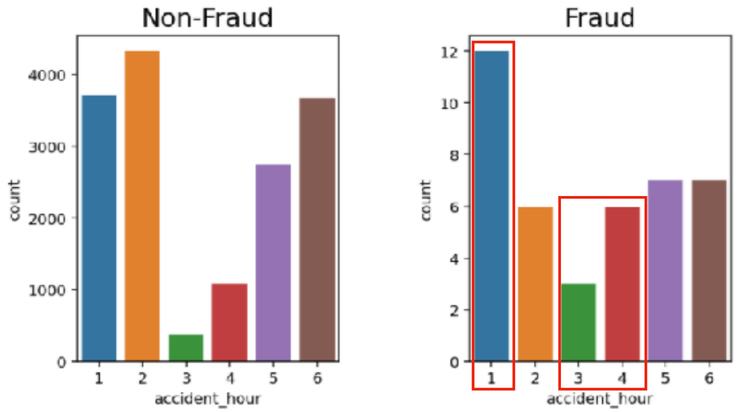
(1부터 차례대로 ‘2~5시간’, ‘6~9시간’, ‘10~36시간’, ‘36시간 초과’, ‘0~1시간’임)

사기군에서는 비교적 단시간 대여 비율은 낮고, 장시간 (10~36시간) 대여 비율은 높게 나타남
(각각 약 14%p, 11%p 차이)

2) 이용/사고 관련 칼럼 EDA 결과

6개 항목 (자기부담금, 이용시작시간, 대여기간, 사고시각, 사고위치, 전면손상 여부)에서 두 그룹간 차이를 보임

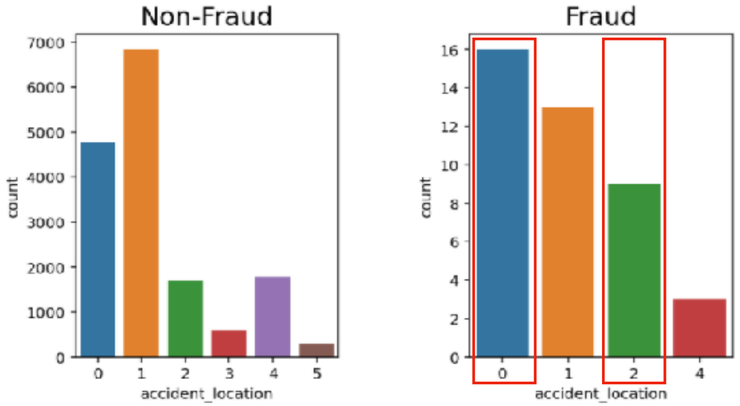
accident_hour



(1부터 차례대로 ‘21~04시’, ‘17~20시’, ‘5~7시’, ‘8~10시’, ‘11~13시’, ‘14~16시’이며, null 대체값인 ‘-1’은 삭제함)

정상군은 퇴근시간 비율이 가장 높고 심야와 오후시간이 그 뒤를 잇는 반면, 사기군은 심야시간 비율이 가장 높고 오전시간도 비교적 높음

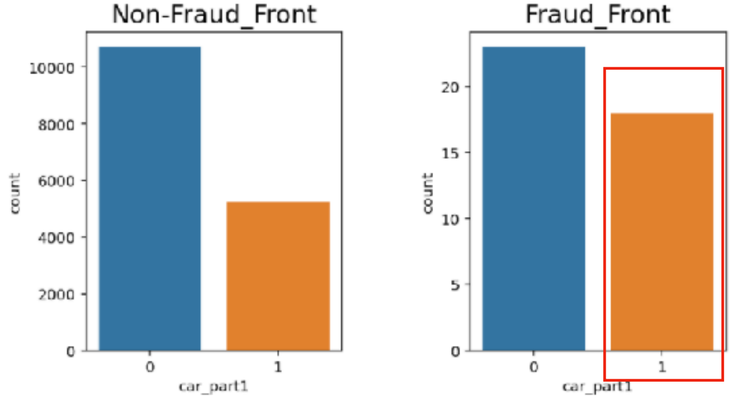
accident_location



(0부터 차례대로 ‘주차장’, ‘일반도로’, ‘이면도로’, ‘고속도로’, ‘쏘카존’, ‘확인불가’임)

정상군은 일반도로 비율이 가장 높은 반면, 사기군은 주차장 비율이 가장 높았으며 이면도로 비율도 비교적 높음

car_part1



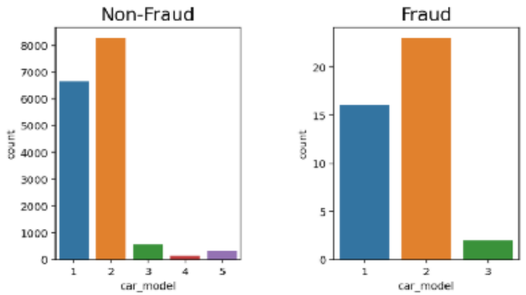
(0이 ‘손상없음’, 1이 ‘손상있음’)

전체 사고건수 대비 전면손상여부 비율은 사기군에서 약 11%p 높게 나타남

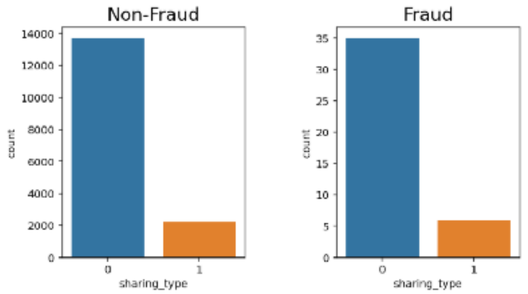
2) 이용/사고 관련 칼럼 EDA 결과

기타 항목들에서는 두 그룹 간 유사한 양상을 보임

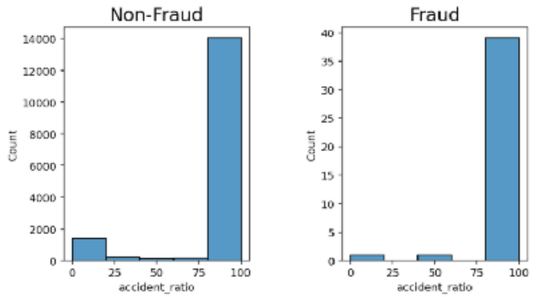
car_model



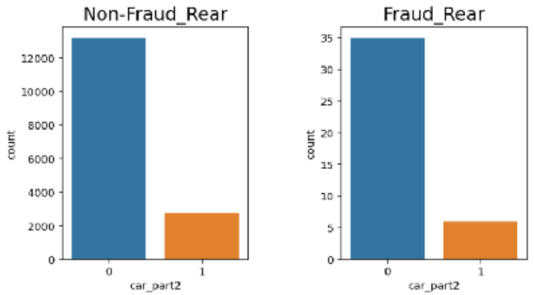
sharing_type



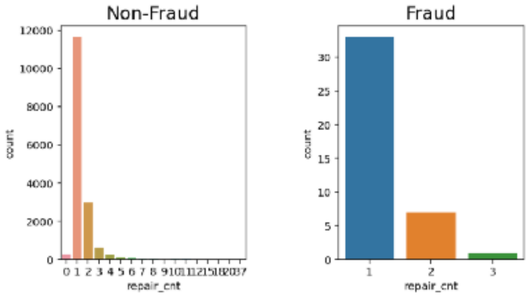
accident_ratio



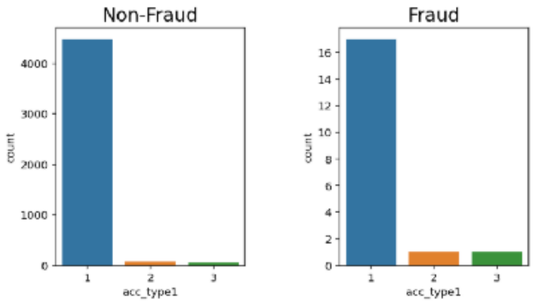
car_part2



repair_cnt



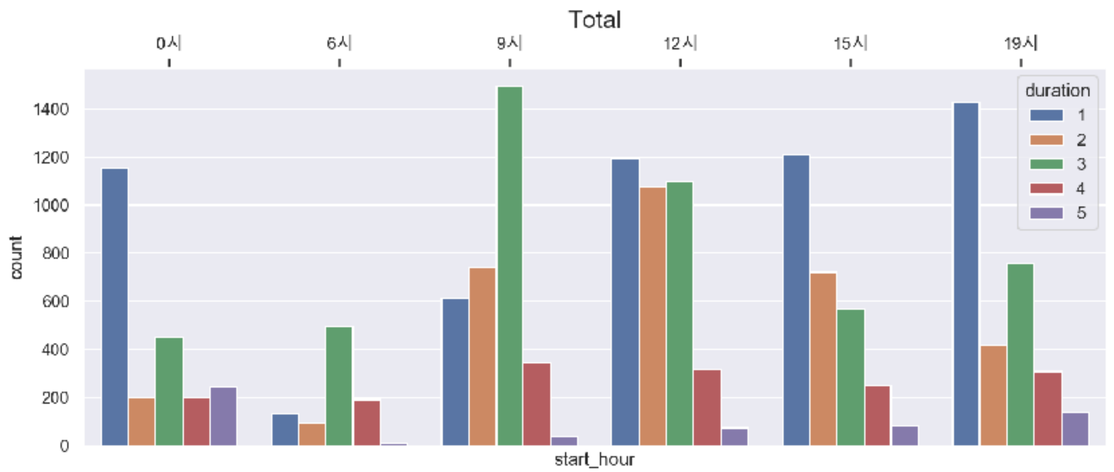
acc_type1



3) 주요 특이사항 공유 (1/3)

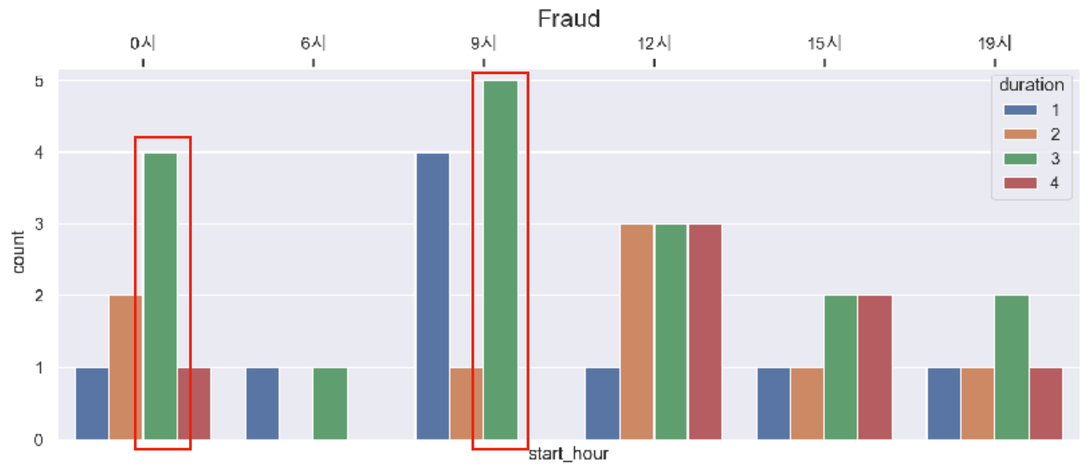
사기군에서는 10시간 이상 장시간 빌리는 비율이 전체 대비 높게 나타남

전체 이용시작시간/대여기간별 분포



오전 시간대를 제외하면, 전반적으로 단시간 (2~5시간) 대여 비율이 높게 나타남

사기군 이용시작시간/대여기간별 분포



전반적으로 장시간 (10~36시간) 대여 비율이 높게 나타나며,
특히 심야 및 오전 시간대에 더욱 두드러지게 나타남

3) 주요 특이사항 공유 (1/3)

저녁/심야시간 및 8시간 이상 대여 시 사용가능한 할인 쿠폰이 하나의 원인으로 추정됨

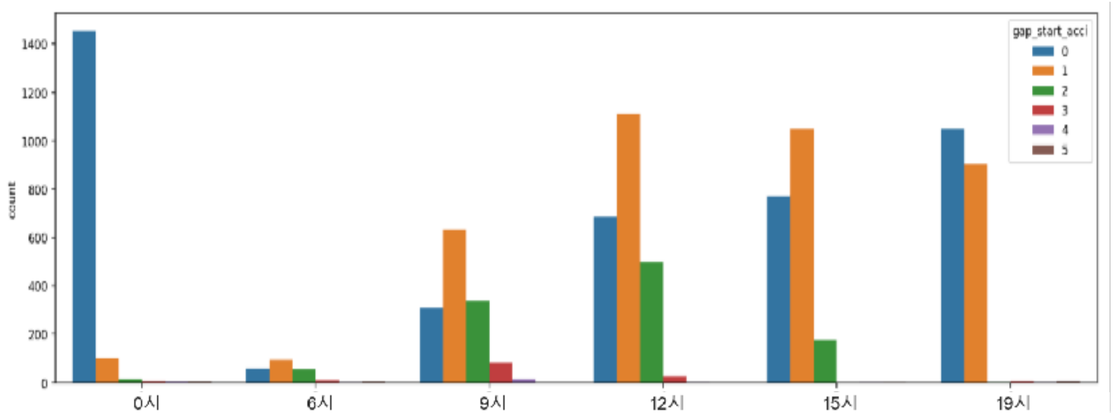
주중 전용 종일 쿠폰 (8~24시간)

주중 전용 저녁/심야 쿠폰 (8~16시간)

3) 주요 특이사항 공유 (2/3)

사기군에서는 이용시작시간 및 대여기간과 무관하게, 이용시작시간과 사고시각 간 시간차가 짧은 것으로 나타남

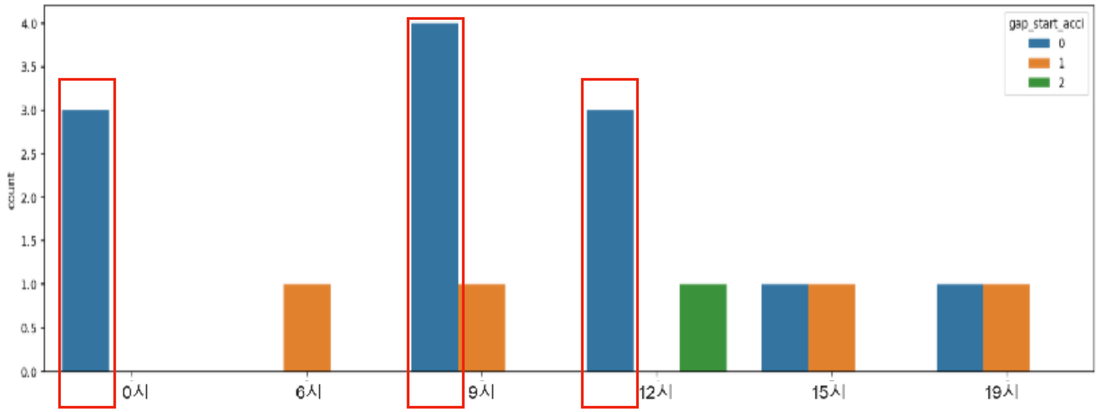
정상군 이용시작시간별 시작시간~사고시각 시간차 분포



(단, 대여기간이 10시간 미만인 데이터에 한정되어 있음)

저녁~심야시간에는 이용시작과 사고시각 간 시간차가 작은 반면,
오전~오후시간에는 비교적 시간차가 커지는 모습을 보임

사기군 이용시작시간별 시작시간~사고시각 시간차 분포



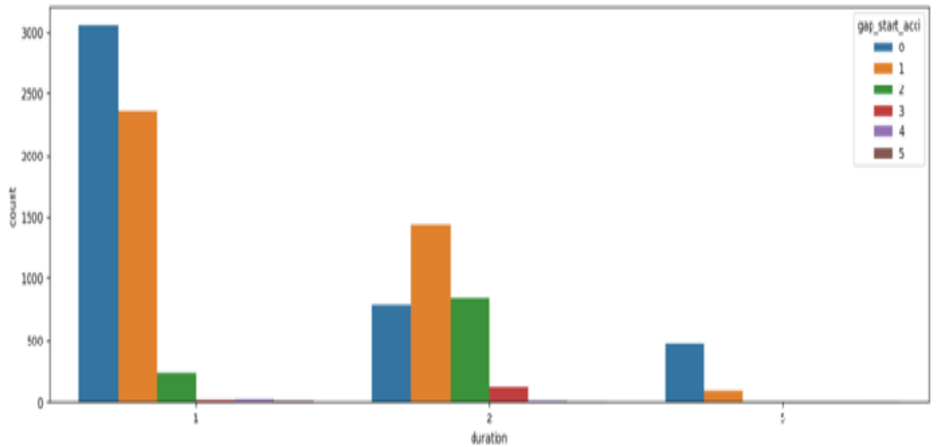
(단, 대여기간이 10시간 미만인 데이터에 한정되어 있음)

이용시작시간과 무관하게 항상 이용시작과 사고시각 간 시간차가
작은 것으로 나타남

3) 주요 특이사항 공유 (2/3)

사기군에서는 이용시작시간 및 대여기간과 무관하게, 이용시작시간과 사고시각 간 시간차가 작은 것으로 나타남

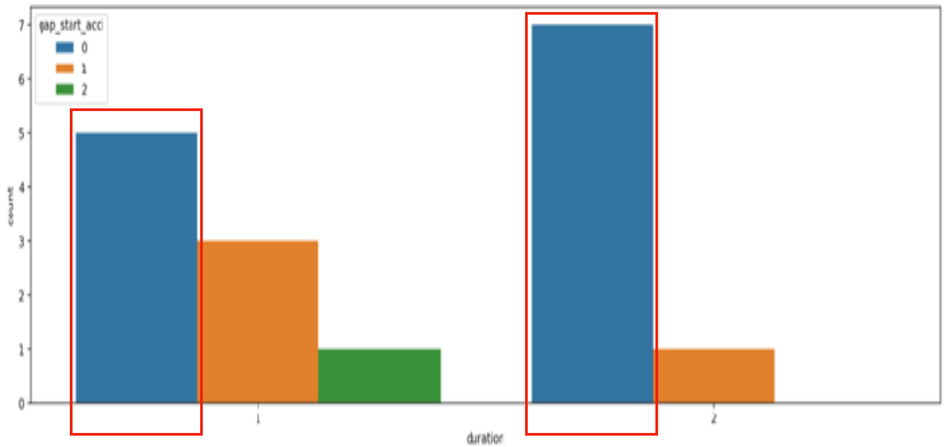
정상군 대여기간별 시작시간~사고시각 시간차 분포



(단, 대여기간이 10시간 미만인 데이터에 한정되어 있음)

대여기간이 길어질수록 이용시작시간과 사고시각 간 시간차도 넓게 분포하는 모습을 보임

사기군 대여기간별 시작시간~사고시각 시간차 분포



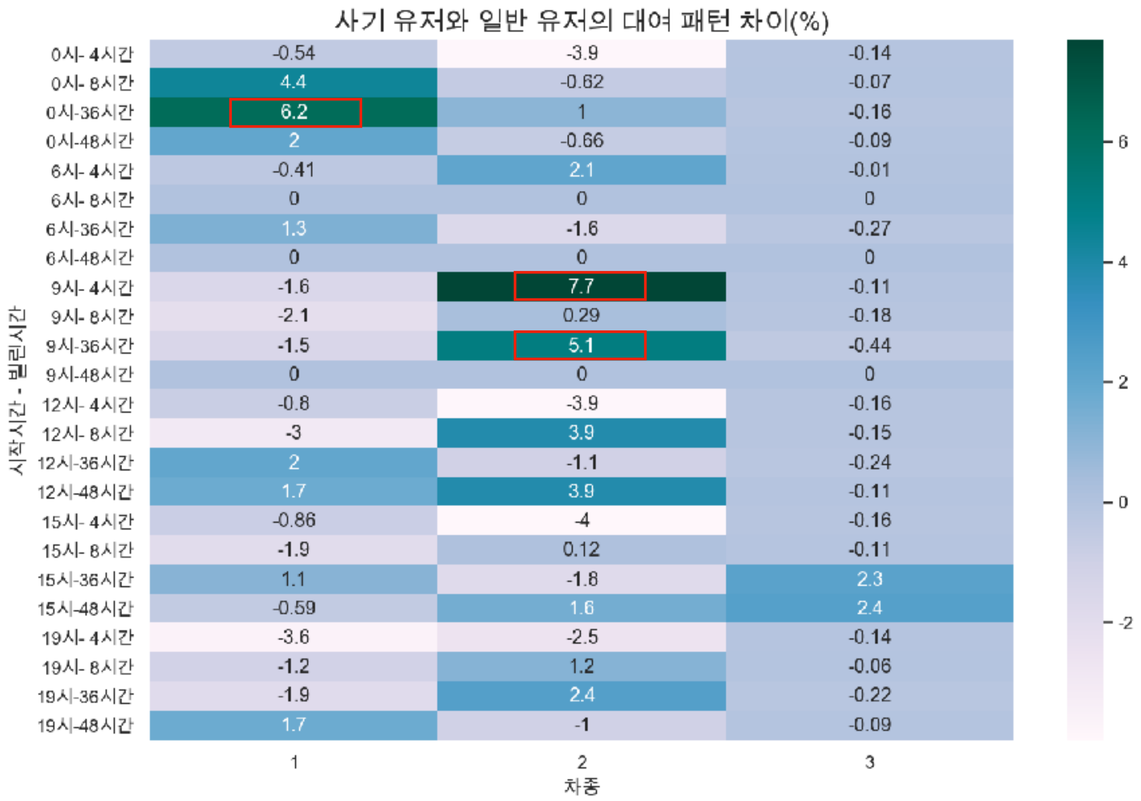
(단, 대여기간이 10시간 미만인 데이터에 한정되어 있음)

대여기간과 무관하게 항상 이용시작과 사고시각 간 시간차가 작은 것으로 나타남

3) 주요 특이사항 공유 (3/3)

사기군에서는 심야에는 소형차를, 오전에는 준중형~중형차를 대여하는 비율이 비교적 높음

차종별 시작시간/대여기간별 정상군~사기군 비율 비교



소형차의 경우, 새벽에 6~36시간 대여가 두드러지는 한편,
준중형~중형차의 경우, 오전에 2~4시간과 10~36시간 대여가
두드러지게 나타남

한계점 및 의문점

한계점

1) 시간 데이터의 한계

- 주중/주말 미구분
- 날짜 데이터가 누락되어 최대 시간 주기를 24시간 이상으로 설정할 수 없는 한계 존재
- 시간 데이터가 이미 범주화되어 있어 데이터 가공에 한계 존재

의문점

1) EDA의 또다른 시각, 관점 여부

- fraud data의 EDA이다 보니, 사기인 데이터와 사기가 아닌 데이터를 비교하는 형태로 EDA를 진행하게 되었음
  → EDA의 측면에서, 데이터를 바라보는 또다른 시각, 관점의 예시가 있었는지가 궁금합니다.

Q & A

E.O.D