교통사고 피해량 예측

-교통사고 피해량이 높은 상황은?

조원: 김현주 박태우 이승후 정윤호 한재현

목차

프로젝트 개요

프로젝트 배경

-교통사고는 매년 많은 인명과 재산 피해를 발생시키는 사회적 문제

-2022년 경상 이상의 피해를 입은 인원은 25만명 가량으로 결코 적은 숫자가 아님

-피해를 최소화하기 위해서는 피해량이 높아지는 특성들을 파악 하고 대비책을 마련하는 것이 필요

프로젝트 목표

-교통사고 피해량과 관련된 **데이터를 수집하고 분석**하여 피해량에 영향을 미치는 **주요 특성을 파악**

-머신러닝 알고리즘을 적용하여 **교통사고 피해량 예측 모델을 개발**

-학습된 모델을 기반으로 교통사고 피해량 예측 결과를 분석하고 해석

-예측 결과를 토대로 **사고 예방 및 관리에 도움이 되는 시사점을 도출하고, 정책 수립 및 대책 마련에 활용할 수 있는 방안을 제시**

모델 개발

데이터 수집

KoROAD 도로교통공단

TAAS 교통사고분석시스템
Traffic Accident Analysis System

1.데이터 제공 기관: 도로교통공단 교통사고분석시스템

-데이터 출처
https://taas.koroad.or.kr/gis/mcm/mcl/initMap.do?menuId=GIS_GMP_AGS_TMM

-데이터 수집 기간: 2020년 1월 1일부터 2022년 12월 31일까지

-데이터 수집 위치: 서울특별시 내 교통사고 발생 지역

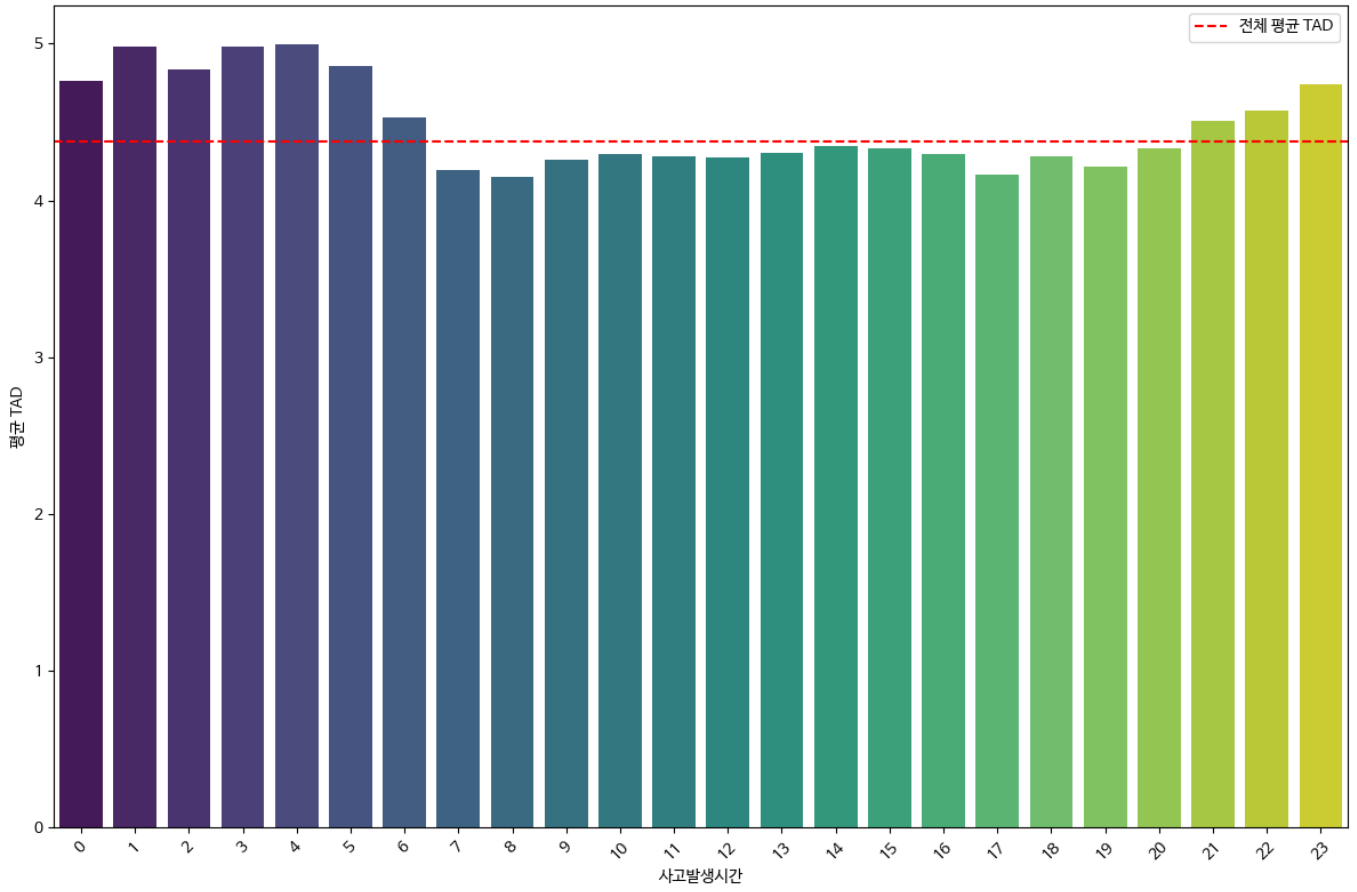
-10만개의 행/22개의 열로 구성

데이터 탐색 – 범주

사고번호	사고일시	요일	시군구	사고내용	사망자수	중상자수	경상자수	부상신고자수	사고유형	법규위반	노면상태	기상상태	도로형태	가해운전자 차종	가해운전자 성별	가해운전자 연령	가해운전자 상해정도	피해운전자 차종	피해운전자 성별	피해운전자 연령	피해운전자 상해정도
2020010100100001	2020년 1월 1일 00시	수요일	서울특별시 양천구 목동	경상사고	0	0	1	0	차대사람 - 기타	안전운전불이행	건조	맑음	단일로 - 기타	이륜	남	56세	상해없음	보행자	남	25세	경상

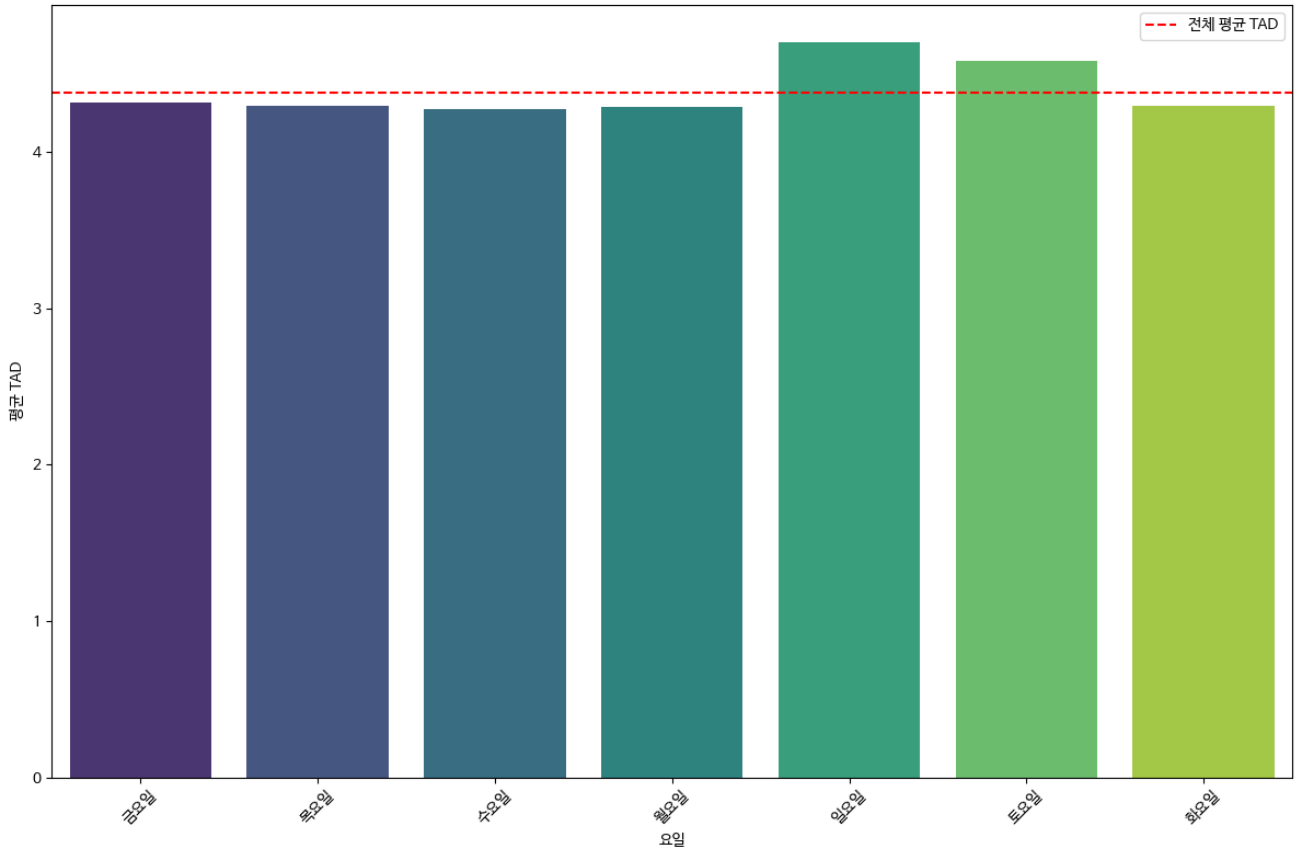
1.성별(2~3): 남, 여, NAN(피해자)
2.사고내용(4): 경상사고, 중상사고, 부상신고사고, 사망사고
3.기상상태(6): 맑음, 흐림, 눈, 비, 안개, 기타
4.상해정도(6~7): 상해 없음, 부상신고, 경상, 중상, 사망, 기타불명, NAN(피해자)
5.노면상태(7) : 건조, 서리/결빙, 적설, 젖음/습기, 침수, 해빙, 기타
6.요일(7): 월~일
7.도로형태(11): 단일로, 교차로, 주차장, 미분류 등등
8.법규위반(11): 안전운전불이행, 중앙선침범, 신호위반, 안전거리미확보 등
9.차종(12~14): 보행자, 이륜, 승용, 승합, 원동기, 자전거, 건설기계, 화물 등
10. 사고유형(17): 차대사람, 차대차, 차량단독 등
11. 시군구(466): 서울시 XX구 XX동

데이터 탐색 – 시각화 -1-



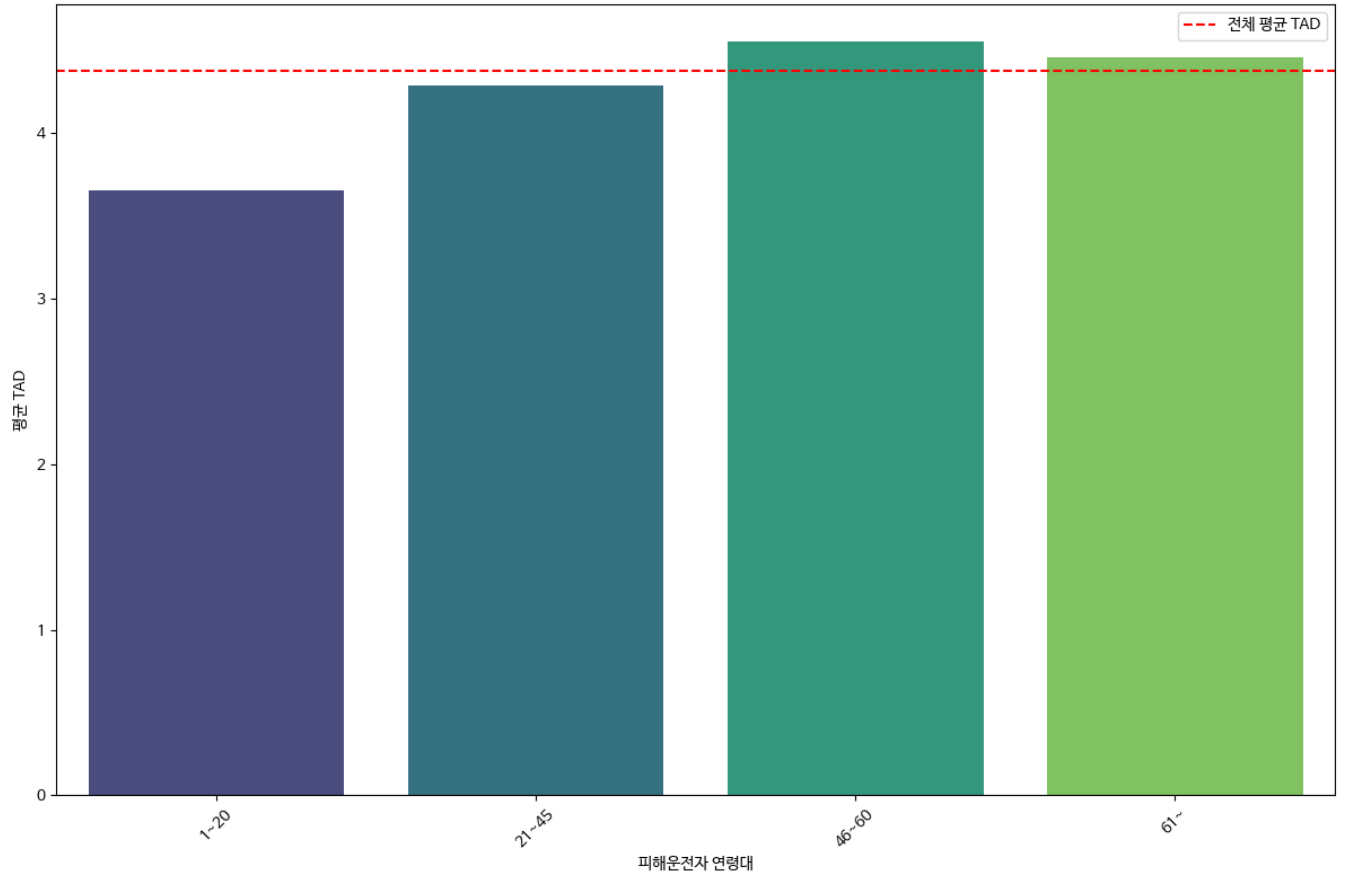
주간 시간대보다 야간 시간대의 교통사고 피해량이 증가

데이터 탐색 – 시각화 -2-



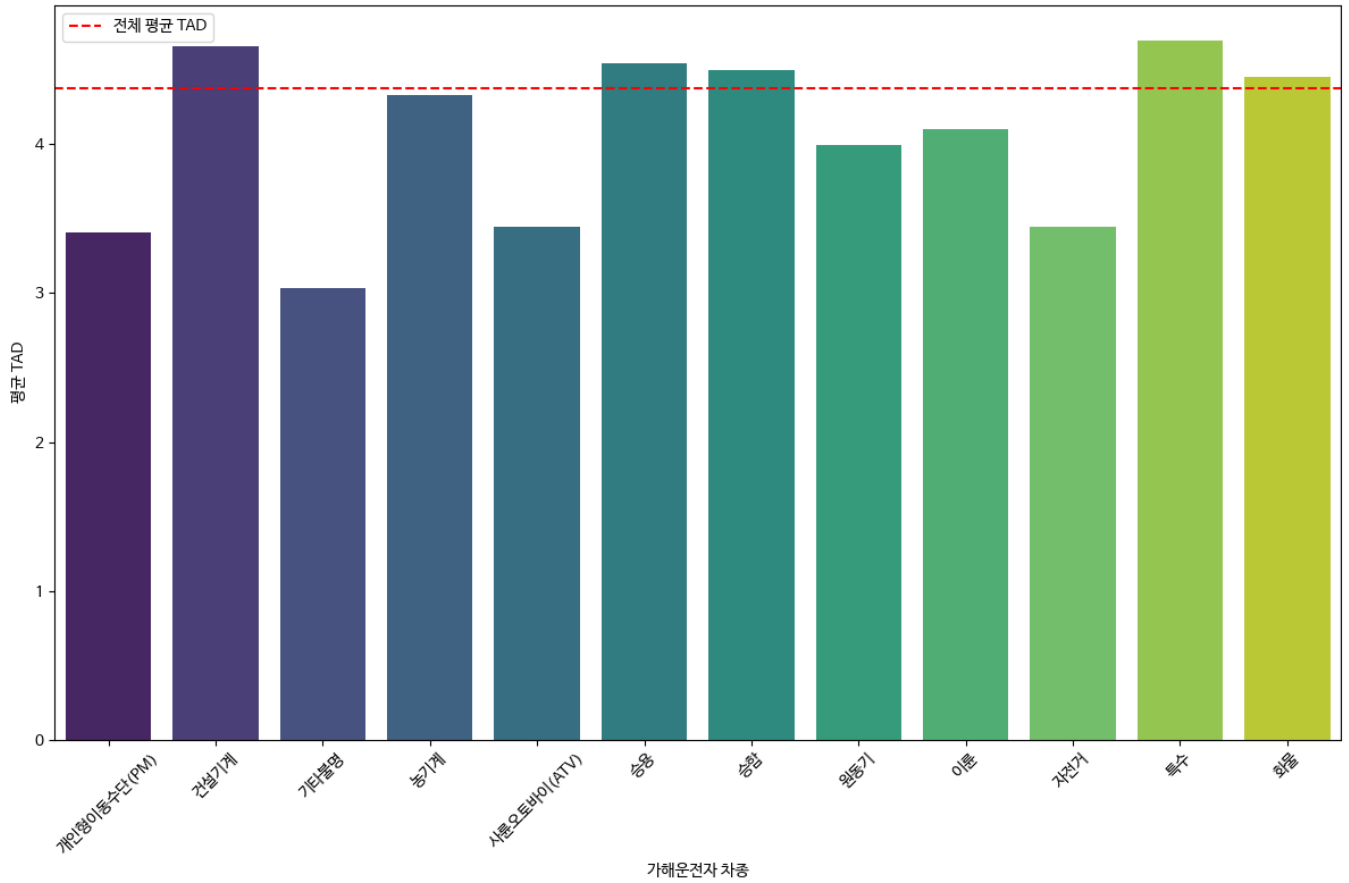
평일보다 주말에 교통사고 피해량이 증가

데이터 탐색 – 시각화 -3-



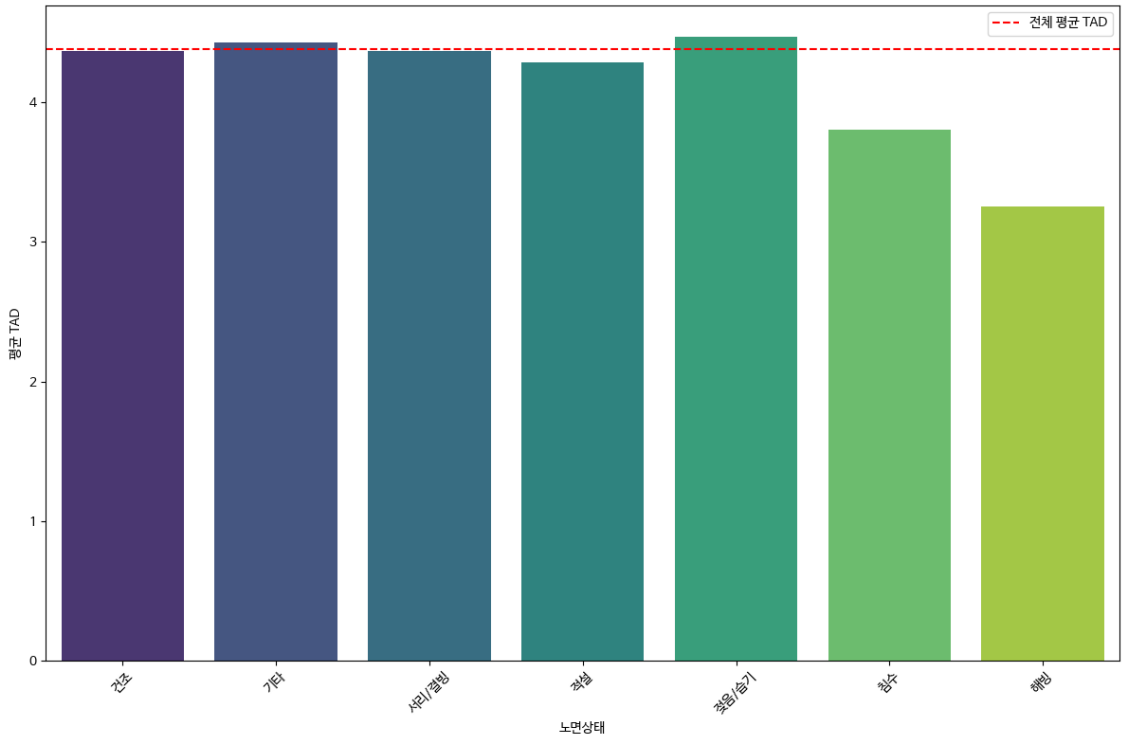
피해자의 연령이 46세를 넘었을때 피해량이 증가

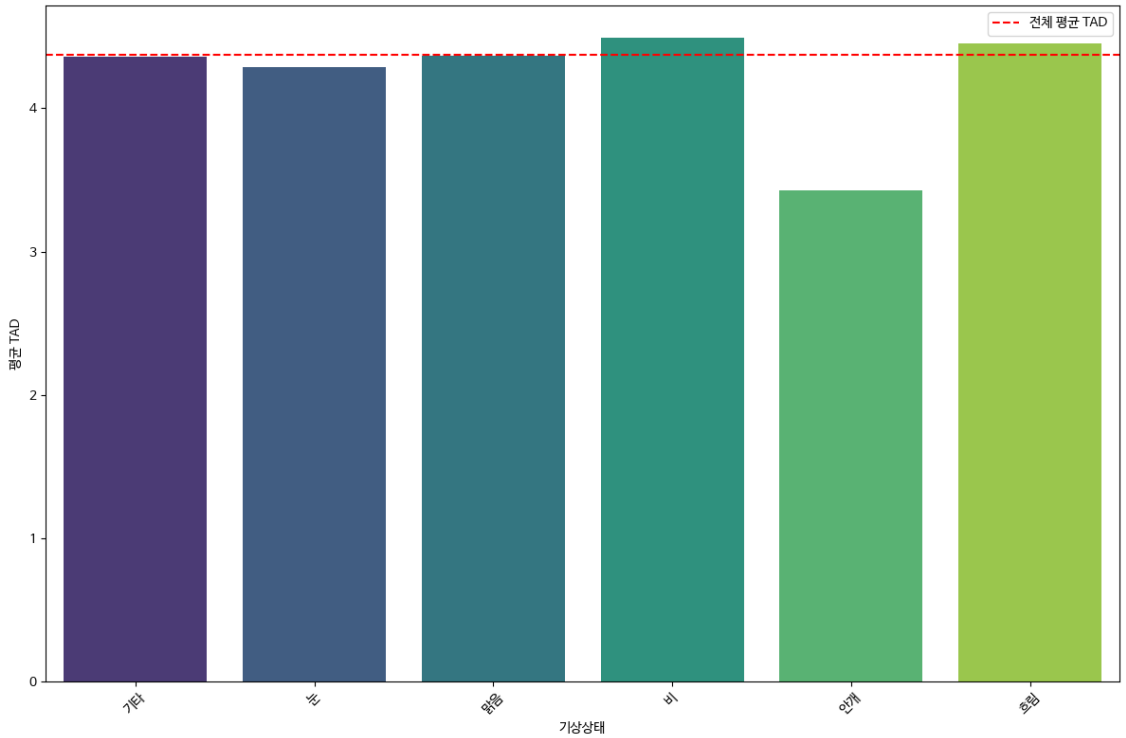
데이터 탐색 – 시각화 -4-



가해운전자의 차종에 따라 피해량 차이 발생

데이터 탐색 – 시각화 -5-





노면/기상상태가 매우 안 좋을때(눈/안개)보다 좋을때(맑음) 피해량이 증가

데이터 전처리

결측치 처리

사고번호	사고일시	요일	시군구	사고내용	사망자수	중상자수	경상자수	부상신고자수	사고유형	법규위반	노면상태	기상상태	도로형태	가해운전자 차종	가해운전자 성별	가해운전자 연령	가해운전자 상해정도	피해운전자 차종	피해운전자 성별	피해운전자 연령	피해운전자 상해정도
2020010100100060	2020년 1월 1일 04시	수요일	서울특별시 성북구 동선동1가	경상사고	0	0	1	0	차량단독 - 전도전복 - 전도	안전운전불이행	건조	맑음	교차로 - 교차로부근	이륜	남	37세	경상				

1. 결측치 발생 -> 차량단독사고의 경우 피해자가 없음

2. 처리 방법
-피해운전자 차종/성별/연령 -> 차량단독사고라는 별도의 범주를 만들어 처리
-피해운전자 상해정도 -> 피해자가 없으므로 상해 없음으로 처리

범주형 데이터 처리 -1-

사고번호	사고일시	요일	시군구	사고내용	사망자수	중상자수	경상자수	부상신고자수	사고유형	법규위반	노면상태	기상상태	도로형태	가해운전자 차종	가해운전자 성별	가해운전자 연령	가해운전자 상해정도	피해운전자 차종	피해운전자 성별	피해운전자 연령	피해운전자 상해정도
2020010100100001	2020년 1월 1일 00시	수요일	서울특별시 양천구 목동	경상사고	0	0	1	0	차대사람 - 기타	안전운전불이행	건조	맑음	단일로 - 기타	이륜	남	56세	상해없음	보행자	남	25세	경상

1. 사고일시 -> datetime으로 변환 ->시간 정보만 남김 ->주간 시간대와 야간 시간대로 구분
2. 요일 -> 평일과 주말로 구분
3. 시군구 -> 범주가 많아 구 정보만 저장(서울특별시 양천구 목동 -> 양천구)
4. 노면/기상상태: 일반(건조/맑음)상황과 위험(비/흐림), 매우위험(눈,안개) 상황으로 나눔
5. 연령: 미성년층/청년층/중장년층/노년층으로 범주화
6. 차종: 보행자(피해자)/소형(오토바이, 자전거 등)/중대형(승용, 화물 등)으로 범주화
7. 이후 범주형 데이터가 있는 특성들은 원-핫 인코딩 적용

범주형 데이터 처리 -2-

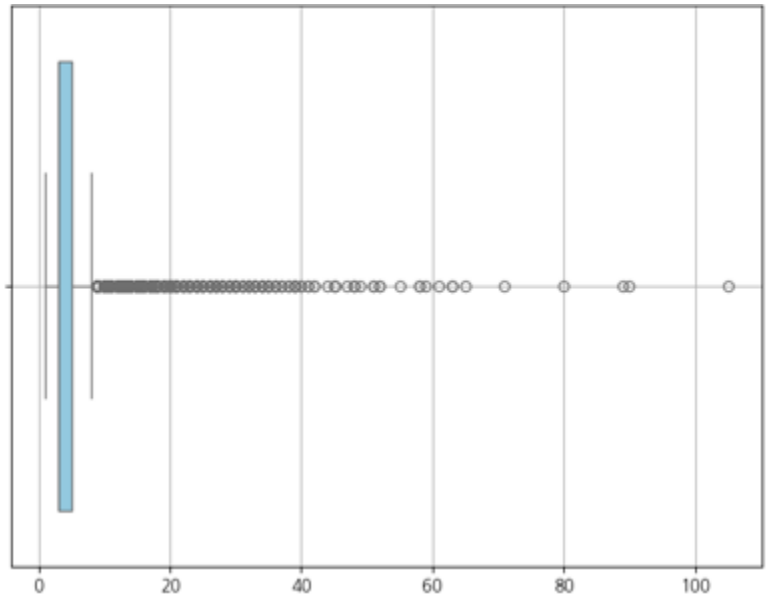
TAD	시군구	평일/주말_주말	평일/주말_평일	시간대_야간	시간대_주간	법규위반_과속	법규위반_교차로운행방법위반	법규위반_기타	법규위반_보행자보호의무위반	법규위반_불법유턴	법규위반_신호위반	법규위반_안전거리미확보	법규위반_안전운전불이행	법규위반_중앙선침범	법규위반_직진우회전진행방해	법규위반_차로위반
3	양천구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
3	노원구	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0
9	용산구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
3	영등포구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
19	구로구	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0
16	강서구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
9	양천구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
12	노원구	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0
3	도봉구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
6	강북구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
3	동대문구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
1	마포구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
3	노원구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
5	서초구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
3	성북구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
6	용산구	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
10	강남구	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
6	중구	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0

피쳐 엔지니어링

1. 종속변수로 쓰일 TAD 특성 생성(부상신고자수, 경상자수, 중상자수,사망자수에 각각 가중치를 주고 생성)

2. 미사용 특성 제거: 사고번호(단순 ID), TAD에 포함된 특성(부상신고자수, 경상자수, 중상자수, 사망자수)

이상치 처리 & 스케일링



종속변수로 쓰일 TAD의 분포를 확인 결과
이상치 처리 필요하다 판단해 Z점수 기반으로 이상치 처리
그 이후 MinMaxScaler로 종속변수 스케일링 진행

데이터 분할

1.X: 원핫 인코딩 이후 특성(85) - 불명확한 특성 제거(ex: 사고유형_차량단독 – 기타)
-시군구 특성(외부데이터 없이 독립적으로 쓰이기 힘듬)

2.Y(TAD_scaled): 스케일링 진행한 교통사고 피해량

3.정확도를 높이기 위해 훈련 70% 테스트 30% 사용/랜덤 시드는 42로 설정

모델 선택 및 튜닝

모델 선택 및 튜닝 -1-

모델	R2 RANK	MSE RANK	MAE RANK
CatBoost	1	1	2
XGBoost	1	1	2
Random Forest	3	3	1
Decision Tree	4	4	2
Linear Regression	5	5	5
Ridge Regression	5	5	7
K-Nearest Neighbors	7	7	6

지표를 고려해 튜닝을 진행할 모델 선정
->CatBoost/XGBoost/Random Forest/Decision Tree

모델 선택 및 평가 –2-

01 기본 XGBoost 모델 성능

테스트 세트 R2 점수: 0.2892
테스트 세트 MSE: 6.3556
테스트 세트 MAE: 1.2120

→

02 이상치 제거+스케일링 이후 XGBoost 모델 성능

테스트 세트 R2 점수: 0.4439
테스트 세트 MSE: 0.01706
테스트 세트 MAE: 0.07956

모델 선택 및 평가 –3-

랜덤서치를 이용해 하이퍼파라미터 최적화 진행

↓

최적화된 파라미터

RF
'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 4,
'max_features': 'sqrt', 'max_depth': 20, 'bootstrap': False

XGB
'subsample': 0.7, 'n_estimators': 300, 'min_child_weight': 1, 'max_depth': 10, 'learning_rate': 0.05, 'gamma': 0.1, 'colsample_bytree': 0.8}

DT
'min_samples_split': 5, 'min_samples_leaf': 7, 'max_features': None, 'max_depth': 9

CB
'learning_rate': 0.05, 'l2_leaf_reg': 9, 'depth': 4

모델 선택 및 평가 –4-

지표	RandomForest	XGBoost	DecisionTree	CatBoost
훈련 세트 R2 점수	1	3	2	4
테스트 세트 R2 점수	2	3	4	1
훈련 세트 MSE	4	3	1	2
테스트 세트 MSE	3	1	2	4
훈련 세트 MAE	4	3	1	2
테스트 세트 MAE	3	1	2	4

테스트 세트 점수가 가장 높은 지표를 가진 XGBoost와 CatBoost를 앙상블

모델 선택 및 평가 –5-

01 기본 XGBoost 모델 성능

테스트 세트 R2 점수: 0.2892
테스트 세트 MSE: 6.3556
테스트 세트 MAE: 1.2120

→

02 이상치 제거+스케일링 이후 XGBoost 모델 성능

테스트 세트 R2 점수: 0.4439
테스트 세트 MSE: 0.01706
테스트 세트 MAE: 0.07956

↓

03 튜닝 이후 XGBoost 모델 성능

테스트 세트 R2 점수: 0.4509
테스트 세트 MSE: 0.01700
테스트 세트 MAE: 0.07952

모델 선택 및 평가 –6-

	XGBoost	CatBoost	Ensemble Voting
훈련 세트 R2 점수	1	3	2
테스트 세트 R2 점수	3	2	1
훈련 세트 MSE	1	3	2
테스트 세트 MSE	3	2	1
훈련 세트 MAE	1	3	2
테스트 세트 MAE	2	3	1

앙상블 이후 테스트 성능이 더 안정적이므로 Ensemble Voting(XGB+CB)를 최종 모델로 선택

모델 선택 및 평가 –7-

01 기본 XGBoost 모델 성능

테스트 세트 R2 점수: 0.2892
테스트 세트 MSE: 6.3556
테스트 세트 MAE: 1.2120

→

02 이상치 제거+스케일링 이후 XGBoost 모델 성능

테스트 세트 R2 점수: 0.4439
테스트 세트 MSE: 0.01706
테스트 세트 MAE: 0.07956

↓

03 튜닝 이후 XGBoost 모델 성능

테스트 세트 R2 점수: 0.4509
테스트 세트 MSE: 0.01700
테스트 세트 MAE: 0.07952

←

04 Ensemble Voting (XGBoost + CatBoost) 모델 성능

테스트 세트 R2 점수: 0.4603
테스트 세트 MSE: 0.01696
테스트 세트 MAE: 0.07945

모델 선택 및 평가 –8-

실제값	예측값	실제값	예측값	실제값	예측값	실제값	예측값	실제값	예측값	실제값	예측값
1	1.072669	3	3.164546	5	5.220693	7	6.405603	9	7.419353	11	9.426682
1	0.966015	3	3.517551	5	5.107208	7	7.134216	9	7.154394	11	8.694615
1	1.3001	3	3.067954	5	5.027007	7	5.32764	9	3.361445	11	8.191296
1	1.099362	3	3.842422	5	5.106604	7	5.585639	9	9.303794	11	8.173205
1	1.054285	3	4.254104	5	5.202935	7	7.407494	9	4.399295	12	8.400766
2	2.967765	4	3.511876	6	5.901167	8	7.999205	10	10.45905	12	8.990215
2	1.990854	4	5.757836	6	7.066812	8	8.885002	10	9.869545	12	8.752133
2	2.314415	4	5.192128	6	4.765941	8	5.284531	10	9.853115	13	11.75799
2	1.533225	4	5.497807	6	4.496521	8	3.610924	10	10.08807	13	12.49178
2	1.682724	4	5.060307	6	4.059387	8	8.214335	10	9.895368	13	10.39382

결과 및 시사점

결과 및 시사점 –1-

결과 도출을 위해 SHAP 라이브러리를 활용 -> 특성 중요도 출력
➔ Voting 앙상블 모델의 직접적인 특성 중요도 출력은 불가능
➔ 개별의 모델(XGBoost/Catboost) 중요도의 평균을 출력

사고유형_차대사람	
길가장자리구역통행중	0.00099079
보도통행중	-4.13E-05
차도통행중	0.02229931
횡단중	0.100108156
사고유형_차대차	
정면충돌	-0.081193277
추돌	-1.765063915
측면충돌	0.535187415
후진중충돌	0.360166179
사고유형_차량단독	
공작물충돌	-0.275170219
도로외이탈 - 추락	0.000398726
전도전복 - 전도	-0.006391187
전도전복 - 전복	-0.014042529
주/정차차량 충돌	0
법규위반	
과속	0.027375067
교차로운행방법위반	-0.124923527
보행자보호의무위반	1.57757102
불법유턴	0.022769483
신호위반	0.680208629
안전거리미확보	0.405823059
안전운전불이행	-0.157861312
중앙선침범	0.02185894
직진우회전진행방해	0.0125167
차로위반	-0.008465182
노면상태_위험도	
건조	0.055304869
위험(적설, 침수)	0.033716282
젖음/습기	0.032863656
기상상태_위험도	
맑음	0.121419692
매우 위험(안개, 눈)	0.076264248
위험(비, 흐림)	0.06845543
도로형태	
교차로 - 교차로부근	0.118009991
교차로 - 교차로안	-0.519096068
교차로 - 교차로횡단보도내	-0.070569232
단일로 - 고가도로위	-0.045727316
단일로 - 교량	-0.265683009
단일로 - 일반	0.035482292
단일로 - 지하차도(도로)내	-0.015505867
단일로 - 터널	-0.027716311
미분류 - 미분류	-0.013045427
주차장 - 주차장	0.032259186

가해운전자 차종_분류	
소형	1.951427016
중대형	0.138857446
가해운전자 성별	
남	0.123691405
여	0.060516421
피해운전자 차종_분류	
보행자	0.759885671
소형	3.037295905
중대형	-0.611720133
피해운전자 성별	
남	4.585856924
여	-2.928322884
평일/주말	
주말	0.136568725
평일	-0.027066013
시간대	
야간	-0.52369182
주간	-0.243787567
가해운전자 연령대	
노년층	-0.011756812
미성년층	0.313364118
중장년층	0.129215839
청년층	-0.020383932
피해운전자 연령대	
노년층	1.073489021
미성년층	2.00221866
중장년층	-0.091663143
청년층	-0.062730494

01 숫자가 높을수록 교통사고피해량과 연관성이 높음

02 +면 피해량이 올라가고 –면 피해량이 내려감

결과 및 시사점 –2-

노면상태_위험도	
건조	0.055304869
위험(적설, 침수)	0.033716282
젖음/습기	0.032863656
기상상태_위험도	
맑음	0.121419692
매우 위험(안개, 눈)	0.076264248
위험(비, 흐림)	0.06845543

→

01 기상 상황이 좋을 때 교통사고 피해량이 높아짐

02 기상 상황이 안 좋을 때는 운전자들이 서로 조심하고 감속 운전을 하기 때문으로 추정

03 이 데이터를 알리는 포스터나 캠페인을 통해 운전자들이 서로 조심하는 분위기 형성 필요

결과 및 시사점 –3-

사고유형_차대차	값
정면충돌	-0.081193
추돌	-1.765064
측면충돌	0.5351874
후진중충돌	0.3601662
도로형태	값
교차로 - 교차로부근	0.11801
교차로 - 교차로안	-0.519096
교차로 - 교차로횡단보도내	-0.070569

01 교차로 부근/측면충돌 피해량이 높음

02 후진중충돌도 높은 피해량인것을 고려할때 교차로부근에서 우선순위를 몰라 발생하는 사고가 많다고 예측

03 운전면허시험 코스에 필수로 교차로를 넣거나
교차로 교통사고를 감소시켜주는 회전교차로 도입등을 검토

결과 및 시사점 –프로젝트 개선 방향-

1.외부 데이터를 도입해 시군구 데이터를 활용할 수 있었다면 더 좋은 결과를 얻을 수 있었을 것
EX)서울시 대중교통 탑승량 데이터를 시군구 데이터를 기준으로 범주화하여 특성으로 추가

2.각 모델의 하이퍼파라미터 자체에 대한 이해도가 높았더라면, 더 나은 튜닝이 가능했을 것

3.프로젝트 진행 속도를 조금 더 높여서 빠르게 모델 선택 과정까지 진행하고, 전처리 과정부터 다시 검토하면서 성능을 비교했다면 개발 효율이 더 좋았을 것

출처

1.데이터
-https://taas.koroad.or.kr/gis/mcm/mcl/initMap.do?menuId=GIS_GMP_AGS_TMM

2.내용
-교통사고 피해자수: http://www.newsmp.com/news/articleView.html?idxno=234147
-회전교차로 도입: https://www.donga.com/news/Society/article/all/20220418/112935040/9

감사합니다!