2023 DACON

대구 교통사고 피해 예측 AI 경진대회

team 삼총사

주관

KATECH
KOREA AUTOMOTIVE TECHNOLOGY INSTITUTE
한국자동차연구원

대구디지털혁신진흥원
Daegu Digital Innovation Promotion Agency

CONTENTS

01

대회 개요

대회 개요

- 기간 : 2023.11.15 ~ 2023.12.11 09:59
- 주제 : 대구 교통사고 피해 예측 – 시공간 정보로부터 사고위험도(ECLO) 예측 AI 모델 개발
- 목적 : 데이터 분석 알고리즘을 통한 교통사고 원인 규명 및 사고율 감소
- 심사 기준: RMSLE(Root Mean Squared Logarithmic Error) of ECLO
  - ECLO = 사망자수 * 10 + 중상자수 * 5 + 경상자수 * 3 + 부상자수 * 1
- 평가 방법 :

항목	심사기준	점수
예측 정확성	리더보드 Private 점수	50
데이터 활용도	다양한 외부 데이터 활용 정도	10
위험요인 분석	교통사고 위험요인 분석	10
인사이트	교통사고에 대한 인사이트	10
해결책	위험요인 및 인사이트와 연계된 해결책	20

- 주최: 산업통상자원부, 대구광역시 / 주관: 한국자동차연구원, 대구디지털혁신진흥원

02

교통사고 피해 예측 모델

데이터 개요

데이터 선택

데이터 전처리

EDA & Feature Engineering

인코딩 & 스케일링

최종 feature 선택

CV 객체 전략

모델링 및 앙상블

데이터 개요

학습 데이터와 Target 값

train 데이터 칼럼

```
0   ID                  39609 non-null  object
1   사고일시              39609 non-null  object
2   요일                 39609 non-null  object
3   기상상태              39609 non-null  object
4   시군구                39609 non-null  object
5   도로형태              39609 non-null  object
6   노면상태              39609 non-null  object
7   사고유형              39609 non-null  object
8   사고유형 - 세부분류    39609 non-null  object
9   법규위반              39609 non-null  object
10  가해운전자 차종        39609 non-null  object
11  가해운전자 성별        39609 non-null  object
12  가해운전자 연령        39609 non-null  object
13  가해운전자 상해정도    39609 non-null  object
14  피해운전자 차종        38618 non-null  object
15  피해운전자 성별        38618 non-null  object
16  피해운전자 연령        38618 non-null  object
17  피해운전자 상해정도    38618 non-null  object
18  사망자수              39609 non-null  int64
19  중상자수              39609 non-null  int64
20  경상자수              39609 non-null  int64
21  부상자수              39609 non-null  int64
22  ECLO                39609 non-null  int64
```

Train only

Target

Train 데이터

1) train.csv
- 총 39,609 개의 대구 교통사고 데이터
- 총 23개의 칼럼

1) countrywide_accident.csv
- 총 602,775 개의 대구를 제외한 전국 교통사고 데이터
- train 데이터와 칼럼 구조 동일

Test 데이터

- 총 10,963 개의 대구 교통사고 데이터
- 총 8개의 칼럼으로 train only 칼럼, target 미포함

Target

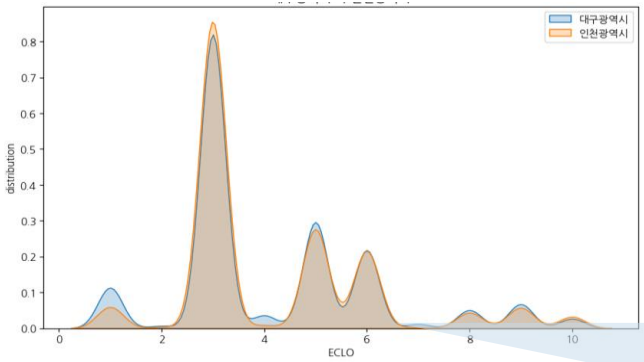
- ECLO
- 사망자*10 + 중상자*5 + 경상자*3 + 부상자*1 (명)

데이터 선택

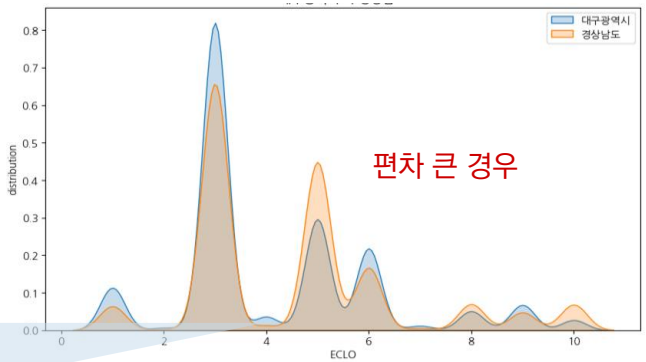
countrywide_accident 에서 총 6개의 대도시 데이터 추가

학습데이터 증가를 위해 전국 데이터에서 '서울, 인천, 광주, 부산, 울산, 대전' 6개 도시 227,699개 데이터를 추가, 대구 포함 총 267,300개의 데이터를 활용

대구광역시 vs 인천광역시 ECLO



대구광역시 vs 경상남도 ECLO



- 대구시와 각 도시별 ECLO 분포의 유사도 (KDE) 분석
- 추가적으로 행정 구역 특성, 도시 규모, 인구 밀도 등을 고려
- 총 17개 도시 중 7개 대도시만을 학습 범위로 설정

데이터 전처리

Test 셋에 없는 Value 삭제

학습 질 향상을 위해 범주형 칼럼인 '기상상태', '도로형태', '노면상태' 의 value값 중 학습데이터에는 있고 test셋에 없는 value 삭제함

countrywide 도로형태

...

교차로 - 교차로횡단보도내	20292
단일로 - 지하차도(도로)내	6782
단일로 - 교량	4179
주차장 - 주차장	3678
단일로 - 터널	2327
단일로 - 고가도로위	1771
미분류 - 미분류	328
단일로 - 철길건널목	12

countrywide 기상상태

맑음	533438
비	40871
흐림	20227
기타	5225
눈	2419
안개	595

countrywide 노면상태

건조	540629
젖음/습기	53234
기타	5450
서리/결빙	2184
적설	1183
침수	67
해빙	27

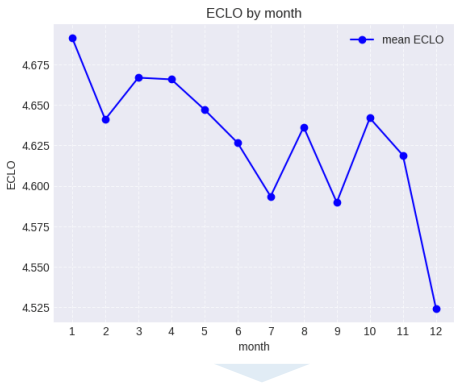
- 도로형태 '철길건널목', 기상상태 '안개', 노면상태 '해빙' value값을 지니는 데이터 삭제

EDA & Feature Engineering

A. 시간 Feature

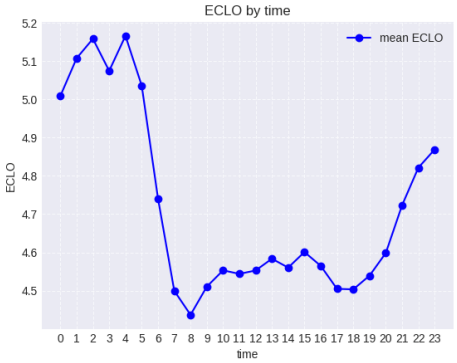
시간 특성별 ECLO 편차 확인, 사고일시 칼럼을 '연', '월', '시간', '요일', '공휴일' 칼럼으로 추가

월별 ECLO



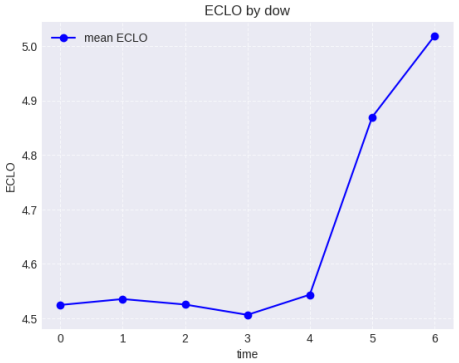
- 최저 : 12월
- 최고 : 1월
- 연말로 갈수록 하락

시간별 ECLO



- 최저 : 오전 6~8
- 최고 : 새벽 0~5
- 활동시간 vs 비활동시간

요일별 ECLO



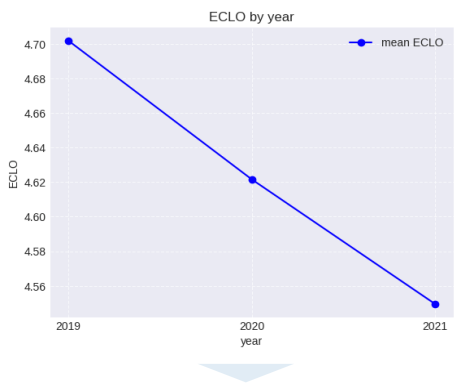
- 최저 : 수요일
- 최고 : 일요일
- 주말동안 높음

EDA & Feature Engineering

A. 시간 Feature

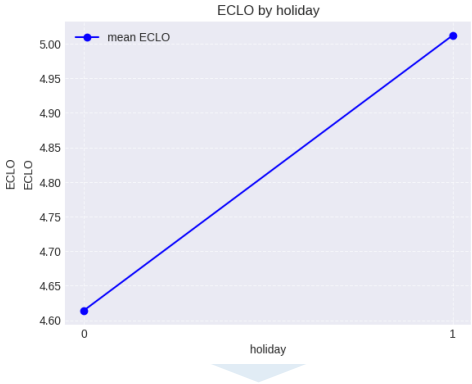
시간 특성별 ECLO 편차 확인, 사고일시 칼럼을 '연', '월', '시간', '요일', '공휴일' 칼럼으로 추가

연별 ECLO



- 최근일수록 낮음

공휴일 ECLO



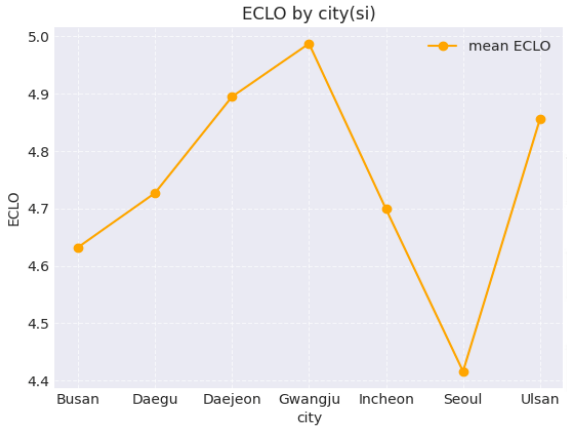
- 공휴일에 높음

EDA & Feature Engineering

B. 공간 Feature

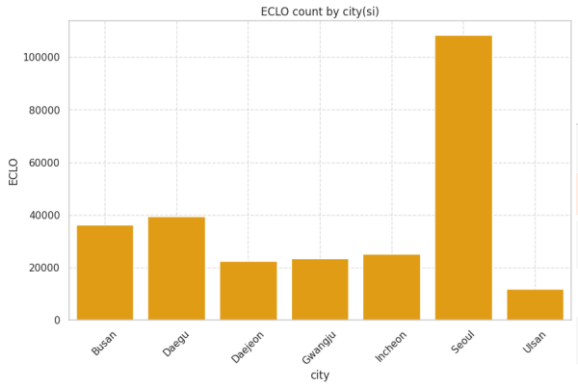
공간별 ECLO 편차 고려, 시군구 칼럼을 분리하여 각각 '시', '구', '동' 칼럼으로 추가

시별 ECLO 평균



	시	ECLO
3	Gwangju	4.986680
2	Daejeon	4.894196
6	Ulsan	4.855915
1	Daegu	4.726547
4	Incheon	4.698550
0	Busan	4.632019
5	Seoul	4.416132

시별 사고횟수



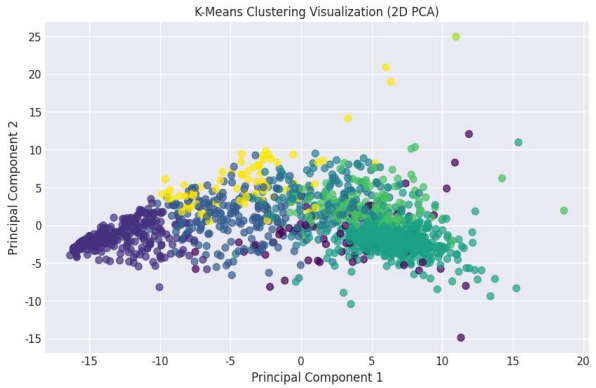
	시	count
5	Seoul	108384
1	Daegu	39601
0	Busan	36241
4	Incheon	25102
3	Gwangju	23423
2	Daejeon	22570
6	Ulsan	11979

EDA & Feature Engineering

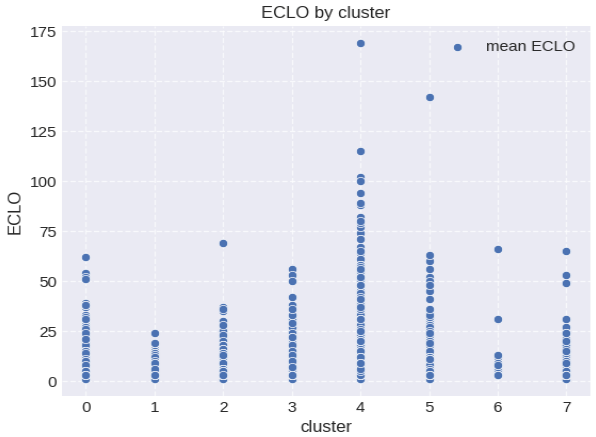
C. Kmeans 군집화 Feature

공간(시-구-동)별 시간별 ECLO 평균을 Kmeans군집화 하여 결과를 군집분석 칼럼으로 추가

Kmeans 군집화



군집별 ECLO 분포



검증 개선(lgbm)

0.421001

▼

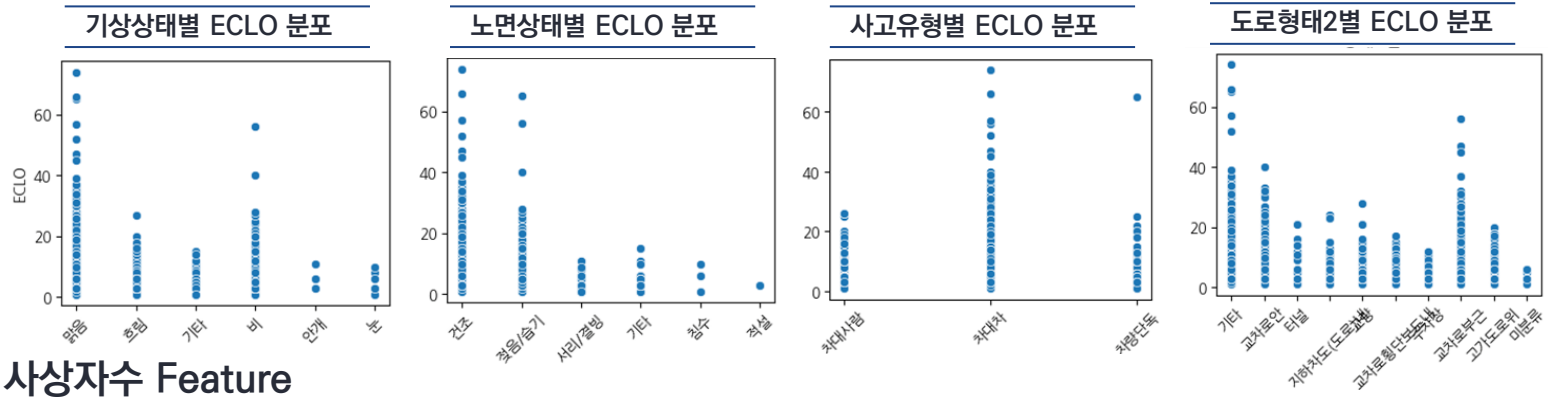
0.420769

- Kmeans 클러스터 n값 중 실루엣 계수 높게 확인된 8개로 군집

EDA & Feature Engineering

D. 기상, 노면, 도로상태 Feature

기상상태, 노면상태, 사고유형, 도로상태 value별 ECLO 차이 확인



E. 사상자수 Feature

사상자 수 (사망자 + 중상자 + 경상자 + 부상자) 의 시간별(시간, 요일) 평균 추가

```
train_eclo['사상자'] = train_eclo['사망자수'] + train_eclo['중상자수'] + train_eclo['경상자수'] + train_eclo['부상자수']

tmp_mean = train_eclo.groupby(['시간','요일'])[['사상자']].mean().reset_index()

train_df = train_df.merge(tmp_mean,how='left',on=['시간','요일'], suffixes= ('','_mean'))
test_df = test_df.merge(tmp_mean,how='left',on=['시간','요일'], suffixes= ('','_mean'))
```

검증 개선(lgbm)

0.421001

▼

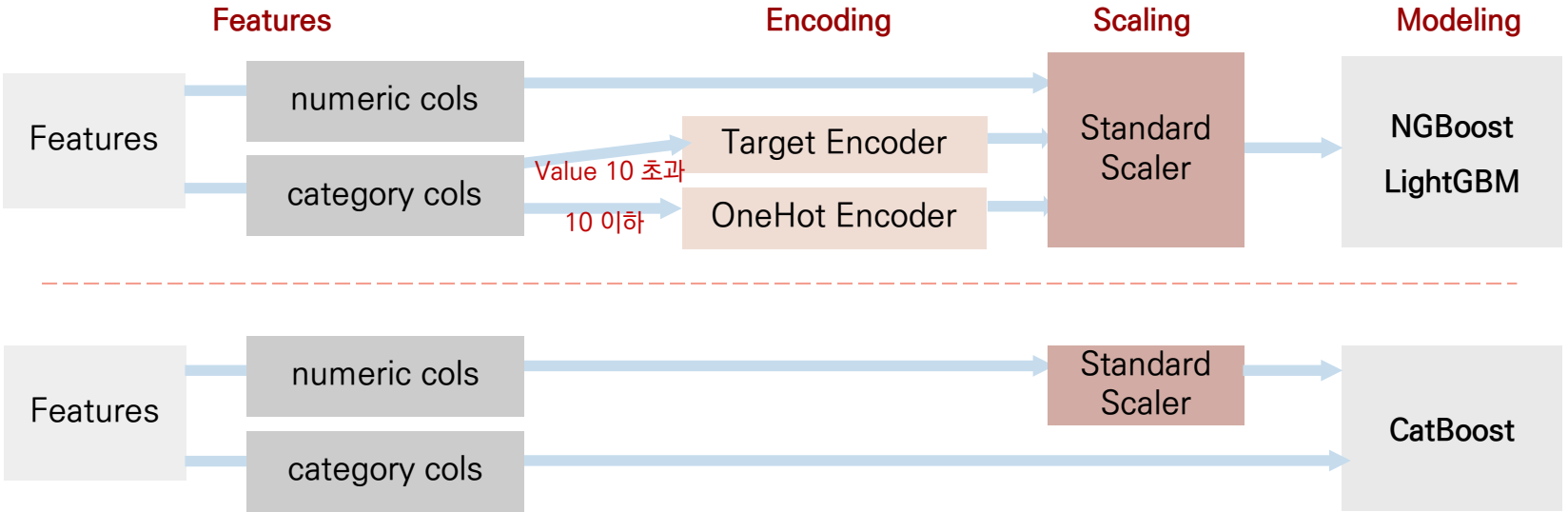
0.420902

인코딩 및 스케일링

B. 인코딩 및 스케일링

일부 칼럼 object 칼럼화한 뒤 numeric, category 및 value 수 기준으로 피쳐 분리하여 인코딩 및 스케일링

- numeric_cols : ['연', '공휴일', '사상자']
- category_cols : ['월','시간','요일','시','구','동','기상상태','도로형태_1','도로형태_2','노면상태','사고유형','군집분석_1']

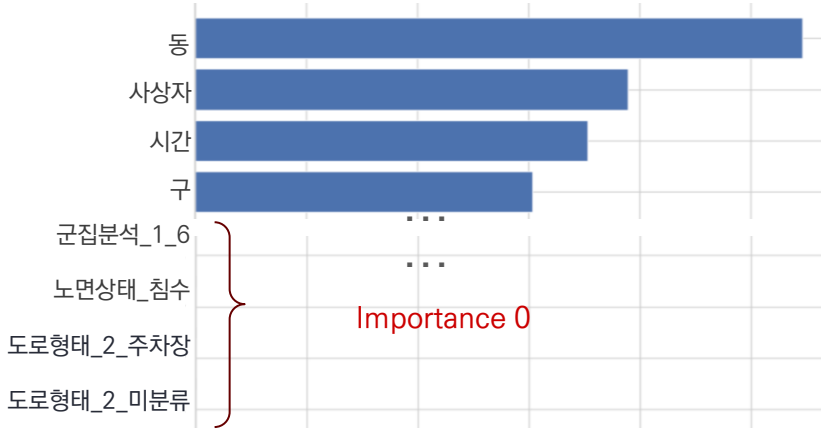


최종 Feature 선택

Feature Importance

인코딩 이후 LGBM, XGB Feature importance 분석 통해 피쳐 드롭 뒤 최종 피쳐 확정

LGBM Feature Importance 그래프



총 4개 칼럼 삭제 : '도로형태_2_미분류', '도로형태_2_주차장', '노면상태_침수', '군집분석_1_6'

최종 Feature

```
0   연                 267300 non-null  float64
1   월                 267300 non-null  float64
2   시간               267300 non-null  float64
3   공휴일             267300 non-null  float64
4   구                 267300 non-null  float64
5   동                 267300 non-null  float64
6   사상자             267300 non-null  float64
7   요일_0             267300 non-null  float64
8   요일_1             267300 non-null  float64
9   요일_2             267300 non-null  float64
10  요일_3             267300 non-null  float64
11  요일_4             267300 non-null  float64
12  요일_5             267300 non-null  float64
13  요일_6             267300 non-null  float64
14  시_광주광역시      267300 non-null  float64
15  시_대구광역시      267300 non-null  float64
16  시_대전광역시      267300 non-null  float64
17  시_부산광역시      267300 non-null  float64
18  시_서울특별시      267300 non-null  float64
19  시_울산광역시      267300 non-null  float64
20  시_인천광역시      267300 non-null  float64
21  기상상태_기타      267300 non-null  float64
22  기상상태_눈        267300 non-null  float64
23  기상상태_맑음      267300 non-null  float64
...
```

CV 객체 전략

A. Stratified KFold 교차검증

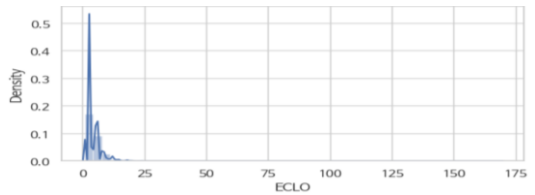
불균형한 타겟에서 클래스 분포를 유지하는 Stratified Kfold 활용하여 성능개선
각 폴드가 타겟 변수 범주의 전반적인 분포를 유지하도록 타겟 범주화

```
# target 균등 분배를 위한 target 범주화
train_df['ECLO_cat'] = 0
train_df.loc[train_df['ECLO'] < 10, "ECLO_cat"] = train_df.loc[train_df['ECLO'] < 10, "ECLO"]
train_df.loc[train_df['ECLO'] >= 10, "ECLO_cat"] = (train_df.loc[train_df['ECLO'] >= 10, "ECLO"]//10)*10
train_df.loc[train_df['ECLO'] >= 70, "ECLO_cat"] = 70
```

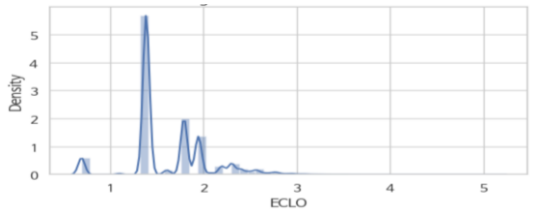
B. Target값 로그변환

ECLO값 로그화를 통해 정규분포화 후 학습하여 성능 개선

Target 분포



Target 로그변환 후 분포



CV 객체 전략

C. 10Fold 학습

전국 데이터 활용으로 학습 데이터 20만개 이상으로 증가하여 10Fold 학습 가능, 일반화 높여 성능 개선

CV 객체 전략별 검증점수 개선 (lgbm)

	Before	After
Target 로그변환	0.4333387	0.421099
10Fold 학습	0.421099	0.421046
Stratified Kfold	0.421046	0.421001
전체 누적적용	0.4333387	0.421001

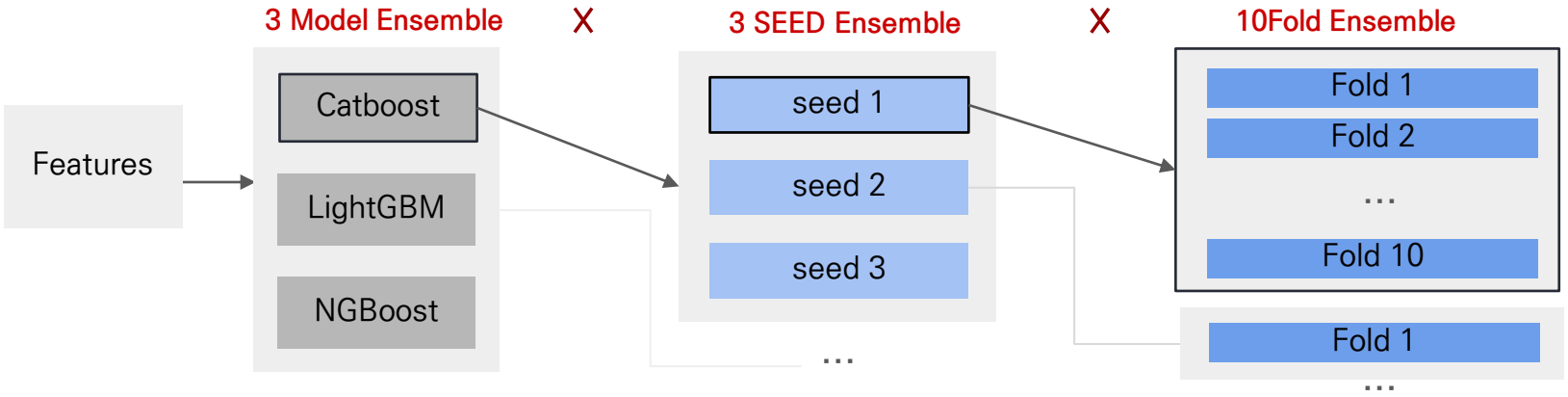
모델링 및 앙상블

Catboost, LightGBM, NGBoost 모델

Catboost, LightGBM, NGBoost 세 모델 사용, 모델별 Optuna 하이퍼파라미터 튜닝

앙상블 : Model – Seed – Fold별 총 90개 모델

모델별, 시드별, 폴드별 앙상블하고 최종 제출은 검증점수 순위 고려하여 가중평균 앙상블



최종 가중평균 : CatBoost 0.4 + LightGBM 0.3 + NGBoost 0.3

03

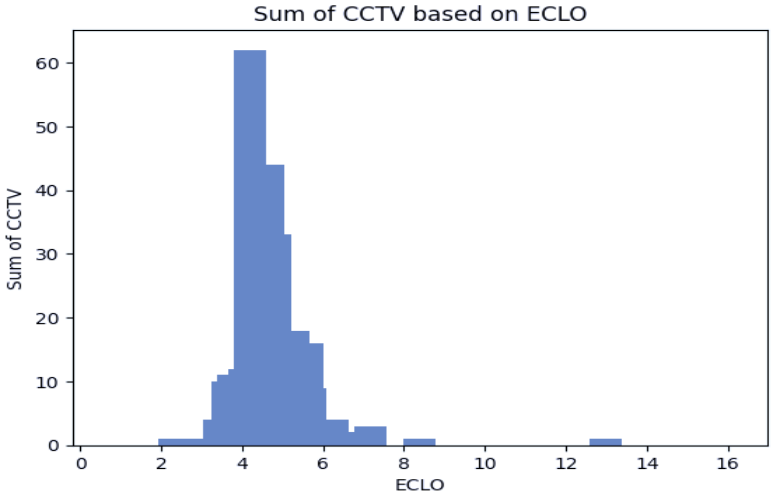
Insight & Solution

인사이트

CCTV 설치의 유용성

대구 CCTV 데이터 분석을 통해, CCTV가 많이 설치된 지역에서 ECLO가 평균적으로 낮은 분포를 보임을 확인

동별 ECLO 평균에 따른 CCTV 총 개수



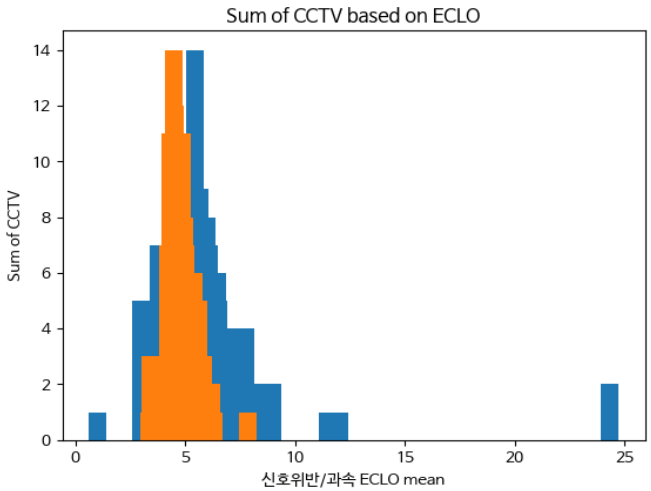
CCTV 설치가 타 지역 대비 효과적일 것이라 예상되는 지역(동)을 확인

해결책

Solution 1 – 신호위반/과속 사고다발구역 CCTV 설치

신호위반/과속 사고인경우 평균 ECLO에 따른 총 CCTV 개수 차이가 전체사고 대비 더 뚜렷
교통사고의 법규위반 유형중 '신호위반', '과속' 최다 동을 신호/과속 사고다발구역으로 지정하여 CCTV 우선 설치

동별 신호위반/과속 사고 ECLO에 따른 CCTV 총 개수



신호위반/과속 사고다발구역 상위 10개 동

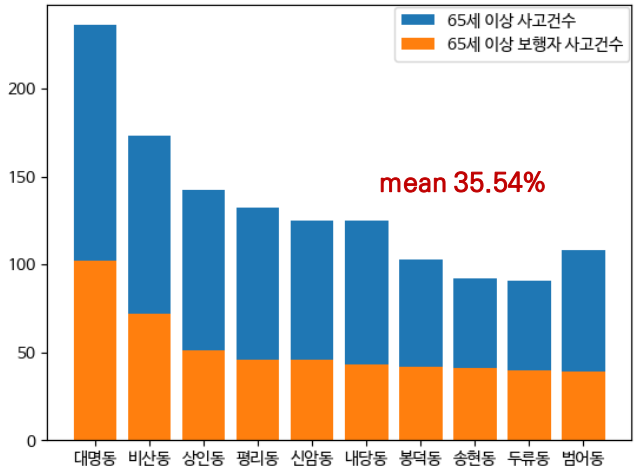
	시_구_동	신호/과속 사고건수	신호/과속 CCTV 설치대수
0	대구광역시 남구 대명동	158.0	27.0
1	대구광역시 북구 침산동	113.0	16.0
2	대구광역시 서구 내당동	104.0	8.0
3	대구광역시 달서구 상인동	99.0	15.0
4	대구광역시 서구 비산동	98.0	14.0
5	대구광역시 서구 평리동	98.0	12.0
6	대구광역시 수성구 범어동	95.0	12.0
7	대구광역시 수성구 만촌동	88.0	15.0
8	대구광역시 달서구 신당동	85.0	7.0
9	대구광역시 동구 신천동	83.0	11.0

해결책

Solution 2 – 노인보호구역 CCTV 설치

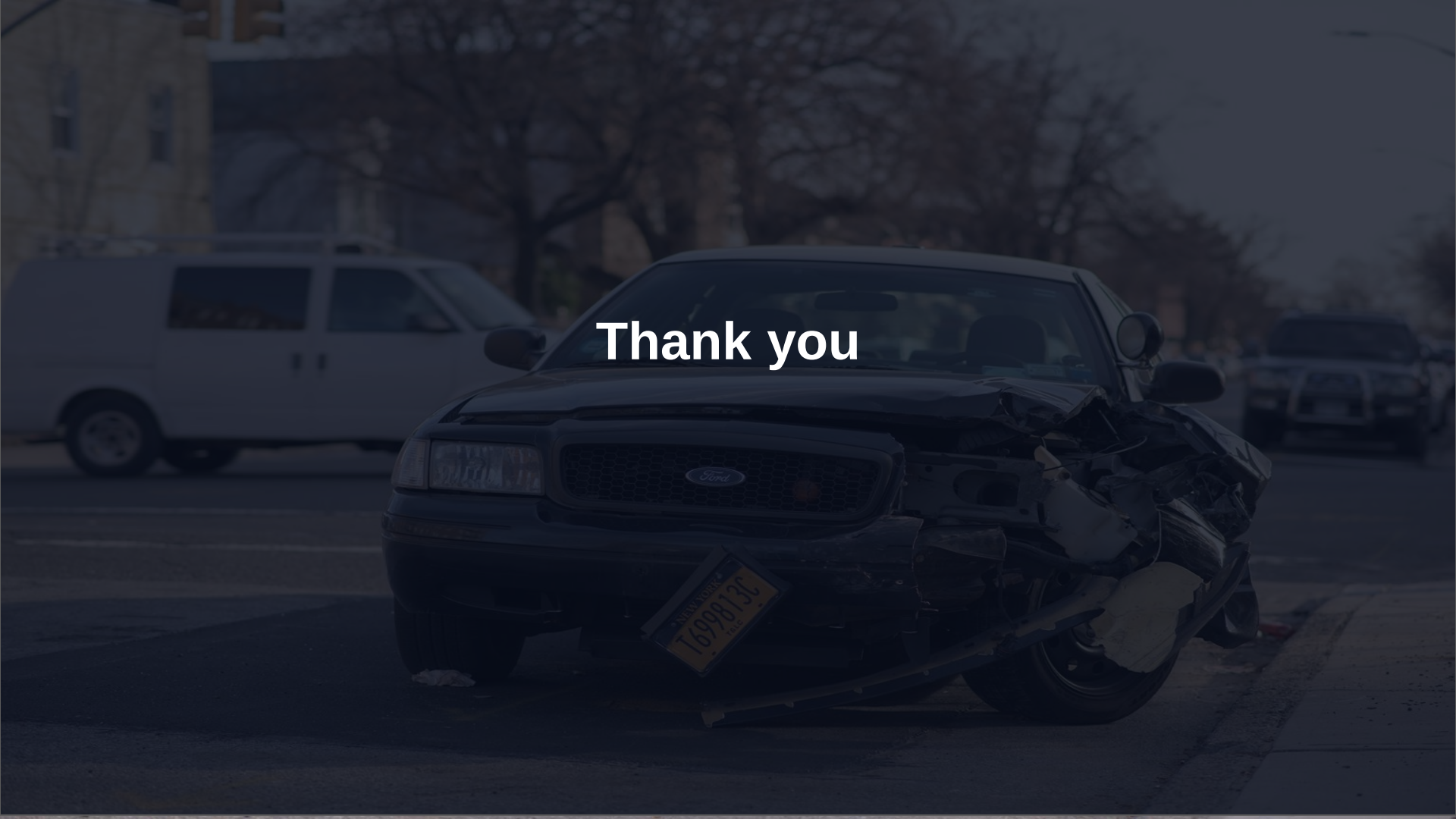
피해자가 65세 이상인 경우, 보행자인 사고 비율이 높음
65세 이상, 보행자 사고 및 '보행자보호의무위반', '신호위반', '과속' 다발 지역에 추가 노인보호구역 지정 및 CCTV 설치

65세 이상인 경우 보행자 비율



노인 보행자 사고다발구역 상위 10개 동

	시_구_동	보행자 사고건수	노인보호구역 수	CCTV 설치대수
0	대구광역시 남구 대명동	102.0	8.0	12.0
1	대구광역시 서구 비산동	72.0	10.0	1.0
2	대구광역시 달서구 상인동	51.0	4.0	3.0
3	대구광역시 서구 평리동	46.0	4.0	0.0
4	대구광역시 동구 신암동	46.0	3.0	0.0
5	대구광역시 서구 내당동	43.0	8.0	5.0
6	대구광역시 남구 봉덕동	42.0	6.0	0.0
7	대구광역시 달서구 송현동	41.0	0.0	0.0
8	대구광역시 달서구 두류동	40.0	12.0	4.0
9	대구광역시 수성구 범어동	39.0	4.0	2.0


Thank you