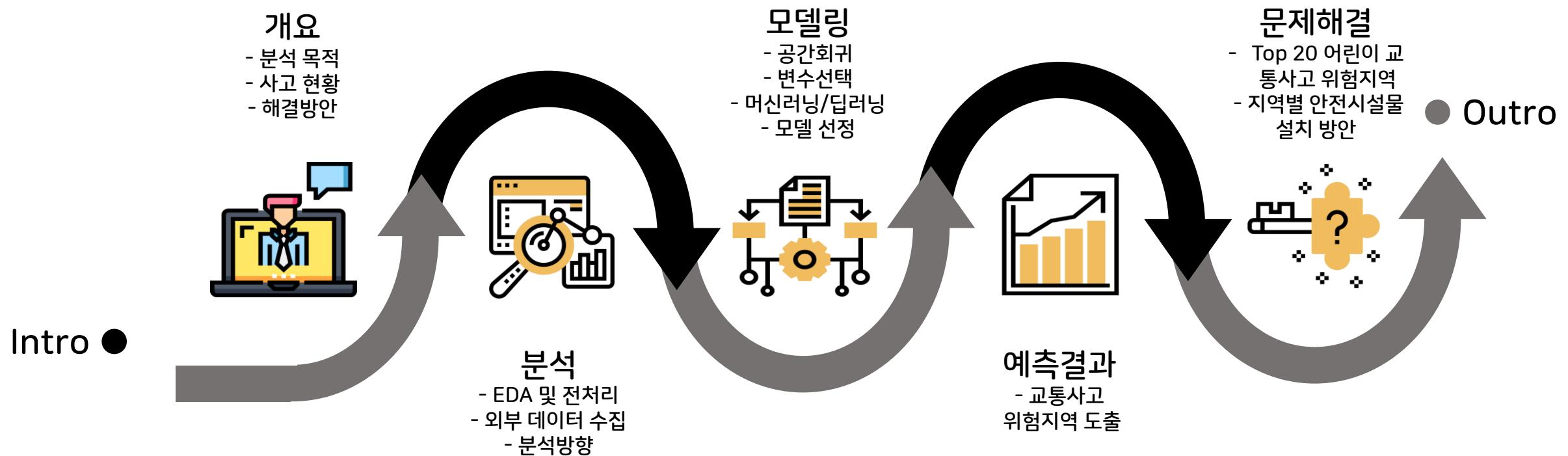




어린이 교통사고 위험지역 도출

Team G.O.C

INDEX



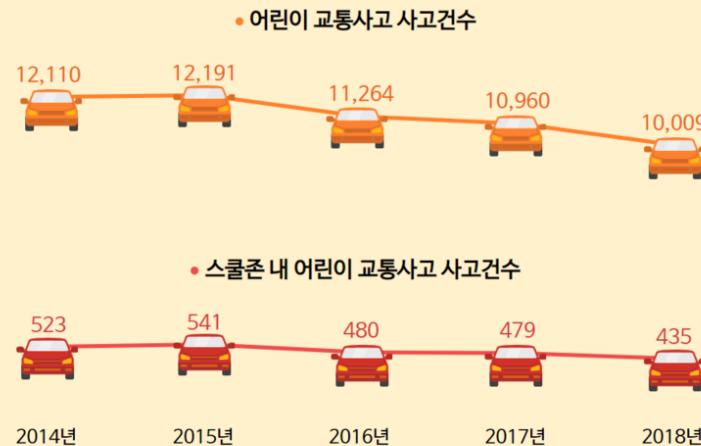
분석 개요

“도로교통공단의 교통사고분석시스템(TAAS) 집계 결과,
최근 3년(2017-2019년) 전국에서 발생한 어린이 교통사고는 3만 2023건”



스쿨존 내 어린이 교통사고 추세

도로교통공단 TAAS 교통사고분석시스템, 단위: 건



■ 어린이 교통사고

11,054건 사고, 28명 사망, 14,115명 부상

전년대비 증감률



■ 스쿨존내 어린이 교통사고

567건 사고, 6명 사망, 589명 부상

전년대비 증감률



출처 : TAAS

어린이 교통사고에 대한 경각심이 높아진 이후에도 어린이 교통사고는 증가 추세를 보이고 있으며,
스쿨존 내 어린이 교통사고는 매년 400건 이상을 유지하고 있다.

오산시 어린이 교통사고 현황

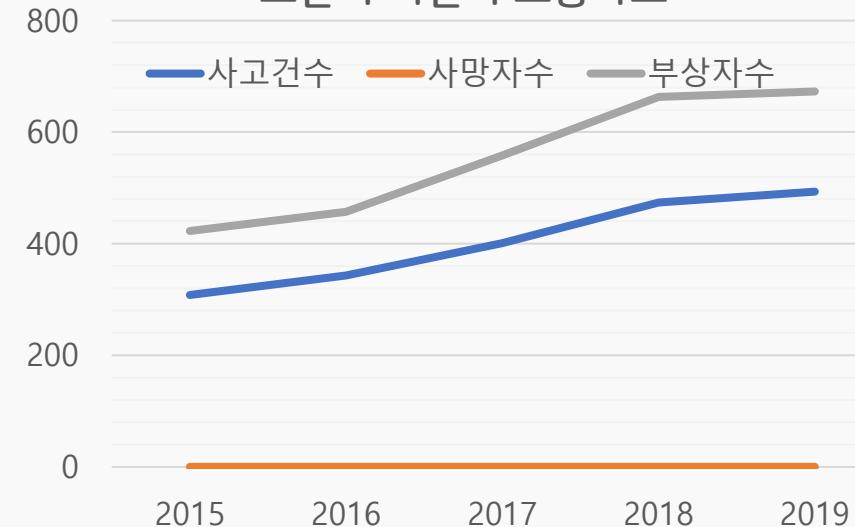
■ 영역별 세부 지표별 산출결과

- 경기 오산시는 운전자 영역(D)에 대한 교통안전도가 가장 취약한 것으로 나타났으며, 운전자 영역 중 신호위반 사고가 취약한 것으로 나타나 해당 세부 사항에 대한 안전대책이 강화되어야 할 것으로 분석됨

영역	점수	등급	순위	세부지표	세부 점수	그룹 평균점수	전체 평균점수
교통약자	80.51	C	18	어린이	76.28	79.10	79.04
				노인	84.73	77.40	78.95
운전자	78.51	D	33	과속	78.01	78.82	78.96
				신호위반	73.49	80.34	78.95
				음주운전	80.74	79.60	78.96
				중앙선침범	81.79	78.53	78.95
도로환경	82.56	B	16	단일로	85.60	78.48	78.95
				교차로	79.52	77.52	78.95

오산시 교통약자부문 교통안전지수가 전국 평균에 비해 낮음

오산시 어린이 교통사고



어린이 교통사고가 매년 증가하는 추세

출처 : TAAS

“오산시의 공간적 특성을 가진 데이터와 교통관련 데이터를 활용하여 어린이 교통사고 위험지역을 도출하고, 기존 어린이 보호구역에 대해 교통사고 감소를 극대화할 수 있는 시설물을 재정비 할 필요가 있음”



EDA / 전처리



데이터 시각화



COMPAS(한국토지주택공사)에서 제공되는 데이터셋의 형태는
MultiPolygon, Points(좌표), LinePath(도로정보)로 나눌 수 있다.

MultiPolygon

- 2. 오산시_어린이교통사고_격자.geojson
- 3. 오산시_차량등록현황_격자.geojson
- 4. 오산시_연령별_거주인구격자(총인구).geojson
- 5. 오산시_연령별_거주인구격자(유소년).geojson
- 6. 오산시_연령별_거주인구격자(생산가능인구).geojson
- 7. 오산시_연령별_거주인구격자(고령).geojson
- 28.오산시_건물연면적_격자.geojson
- 11.오산시_초등학교_통학구.geojson
- 12.오산시_중학교_학군.geojson
- 17.오산시_횡단보도.geojson
- 27.오산시_도로명주소_건물.geojson
- 31.오산시_법정경계(시군구).geojson
- 32.오산시_행정경계(읍면동).geojson
- 33.오산시_법정경계(읍면동).geojson
- 34.오산시_지적도.geojson

Points

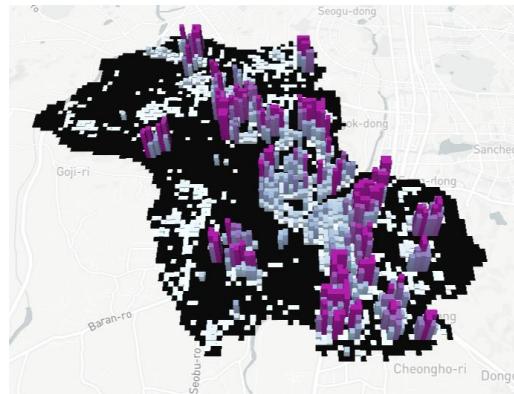
- 1. 오산시_주정차단속(2018~2020).csv
- 8. 오산시_유동인구(2019).csv
- 9. 오산시_어린이보호구역.csv
- 10.오산시_학교위치정보.csv
- 13.오산시_어린이집_유치원현황.csv
- 15.오산시_무인교통단속카메라.csv
- 16.오산시_도로안전표지표준데이터.csv
- 18.오산시_과속방지턱표준데이터.csv -> 외부데이터 사용
- 19.오산시_신호등.geojson
- 20.오산시_CCTV설치현황.csv
- 22.오산시_버스정류장.csv
- 29.오산시_체육시설현황.csv
- 30.오산시_학원 및 교습소 현황.csv

LineString

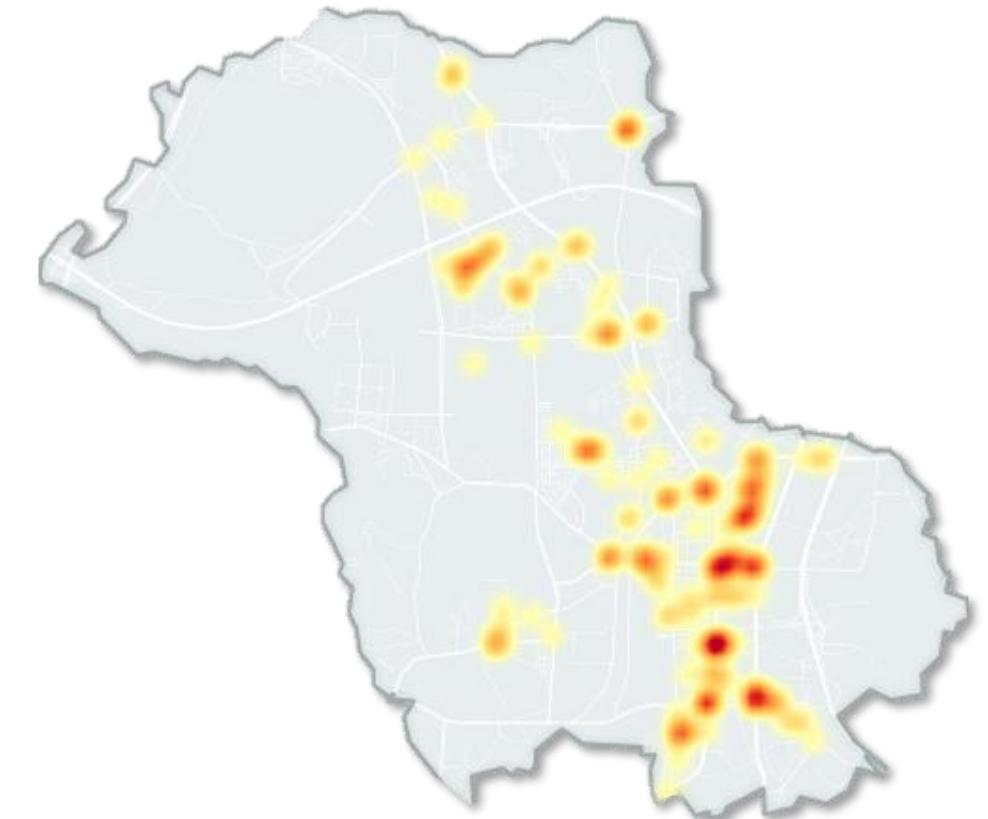
- 21.오산시_인도.geojson
- 23.오산시_상세도로망_LV6.geojson

주요 데이터 시각화 - (1)

4.오산시_연령별_거주인구격자(총인구).geojson



2.오산시_어린이교통사고_격자.geojson

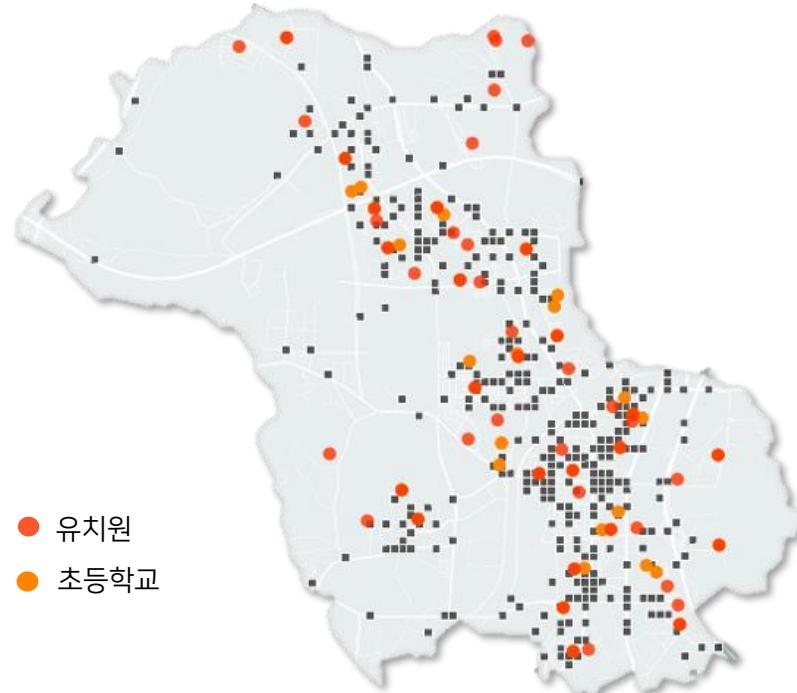


대원동과 신장동에 "오산시 전체 인구의 약 54%가 밀집"

중앙동, 대원동, 신장동에 걸쳐 집중적으로 사고가 분포

주요 데이터 시각화 - (2)

학교, 유치원 위치와 어린이 교통사고 지점



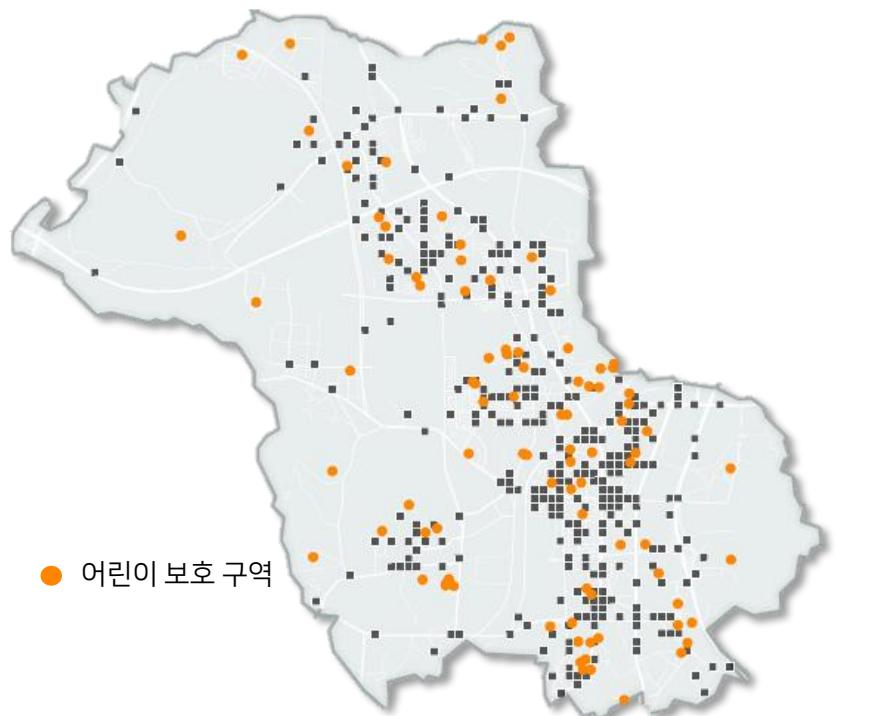
횡단보도와 사고 지점 비교



어린이 교통사고가 일어나는 지점의 분포와 초등학교, 유치원의 위치의 분포가 유사하고 대체로 횡단보도가 있는 도로에서 사고가 남을 대략적으로 알 수 있음

주요 데이터 시각화 - (3)

어린이 보호 구역과 사고 지점 비교



어린이 보호 구역 외에도 사고가 많이 발생하므로,
추가적인 어린이 교통사고 위험지역 선정이 필요함



“시사점 도출”

거주 인구, 학교, 유치원의 위치, 횡단보도 등
다양한 요인이 어린이 교통사고의 발생에
영향을 미칠 것으로 보임



전처리



외부 데이터 수집 - 과속방지턱

* 제공받은 과속방지턱 데이터의 관측치가 너무 적으로 수집에 오류가 있다고 판단해 외부데이터를 수집

사고지점 반경 주요시설 (카카오 지도/로컬 API 활용)	학원_50m이내 어린이 교통사고 지점의 중심좌표 50m 이내에 있는 학원의 수 ※ 키워드와 카테고리, 반지름 parameter를 활용하여 학원의 수를 구함
과속방지턱 (공공데이터포털)	과속방지턱_개수 어린이 교통사고가 발생한 지점(격자)에 속하는 과속방지턱의 개수

내부 데이터 전처리 - Point type

오산시_어린이보호구역

1. 하나의 Point 좌표에 버퍼를 주어 반경 300m의 MultiPolygon 형태의 데이터로 변환

: 위, 경도를 미터법으로 변환 ※위도마다 거리 계산법이 다르고, 위도의 1도와 경도의 1도도 거리 계산법이 다름

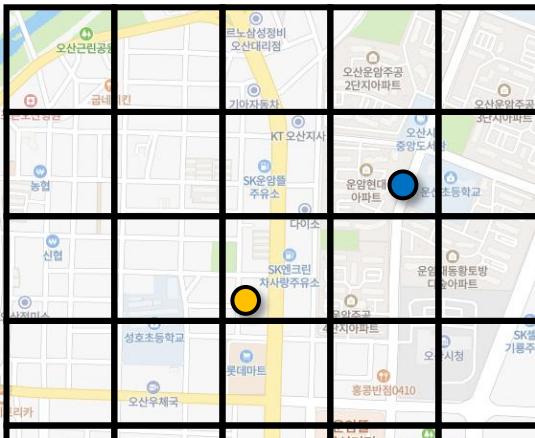
위경도 37, 127도 부근에서 위도와 경도의 1도를 평균하여 약 100km로 설정

따라서 위도와 경도의 1도는 100km이고 [0.001도 = 100m] 를 buffer에 적용

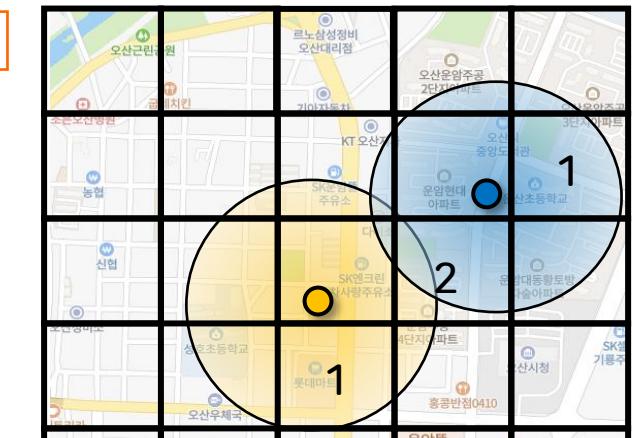
2. 4510개의 격자별로 버퍼를 준 데이터의 MultiPolygon값이 몇 번 겹치는 측정

3. '오산시_어린이교통사고_격자' 데이터셋과 격자를 기준으로 하여 join함

원본 데이터



스쿨존



내부 데이터 전처리 - Point type

오산시_학교위치정보, 오산시_유치원현황

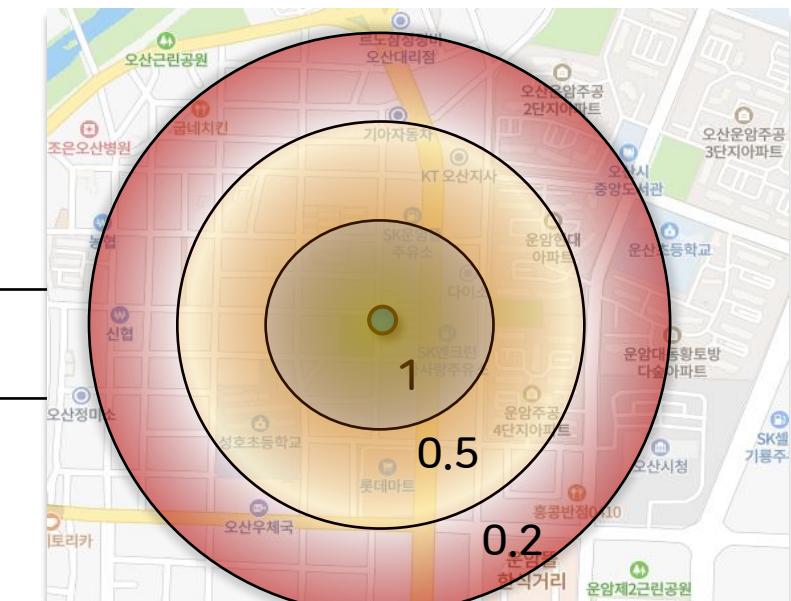
1. 하나의 Point 좌표에 버퍼를 주어 반경 100m/300m/500m의 MultiPolygon 형태의 데이터로 변환

: 위, 경도를 미터법으로 변환

2. 4510개의 격자별로 버퍼를 준 데이터의 MultiPolygon값이 **가중치를 두어** 몇 번 겹치는 측정

: 학교와 가까운 지점일수록 영향치를 높이기 위해 거리별로 가중치를 두기로 함

가중치	반경
1	해당 격자가 초등학교 반경 100m안에 포함될 경우
0.5	해당 격자가 초등학교 반경 300m안에 포함될 경우
0.2	해당 격자가 초등학교 반경 500m안에 포함될 경우



3. '오산시_어린이교통사고_격자' 데이터셋과 격자를 기준으로 하여 **join**함

내부 데이터 전처리 - Point type

오산시_횡단보도

1. 4510개의 격자별로 포함하는 횡단보도 면적 측정

crosswalk_size - 각 격자와 겹쳐지는 횡단보도의 전체(횡단보도 전체) 면적 계산

※횡단보도가 해당격자의 범위를 넘어가더라도 횡단보도의 크기를 계산하기 위해 격자가 포함하는 횡단보도의 전체 면적의 합을 계산

crosswalk_area - 각 격자와 겹쳐지는 횡단보도만의(횡단보도 일부) 면적 계산

※횡단보도의 전체 면적 중 해당 격자에 포함하는 범위까지의 넓이만 계산

2. '오산시_어린이교통_격자' 데이터셋과 격자를 기준으로 하여 **join**함

오산시_신호등, 주정차단속, 단속카메라, 도로안내표지, CCTV

1. '오산시_어린이교통사고_격자' 데이터셋과 격자를 기준으로 하여 **join**함

2. 격자를 기준으로 groupby_sum하여 각 격자가 가지고 있는 신호등 개수를 count

내부 데이터 전처리 - Point type

오산시_도로명주소_건물

1. 4510개의 격자별로 포함하는 건물(전체 / 주거용 / 사무용) 개수 측정

building - 한 격자가 포함하는 건물의 개수

office - 코드정의서에 기입된 건물용도에 따라 사무용으로 분류된 건물의 개수

※건물용도코드가 04401~04499, 10000, 10101~10299에 해당하는 건물

house - 코드정의서에 기입된 건물용도에 따라 주거용으로 분류된 건물의 개수

※건물용도코드가 01000~02007에 해당하는 건물

2. '오산시_어린이교통사고_격자' 데이터셋과 격자를 기준으로 하여 **join**함

내부 데이터 전처리 - Point type

오산시_버스정류장, 오산시_체육시설현황

1. '오산시_버스정류장.csv', '오산시_체육시설현황.csv'를 geojson으로 변환

2. '오산시_어린이교통_격자' 데이터셋과 격자를 기준으로 하여 **join**함

: MultyPoligon과 points 형태의 데이터를 join할 시 특정 격자는 두개 이상의 row를 가짐

Ex)

gid	정류장명	정류장ID
다사45067	오산시티자이1.2단지	223000095.0
다사45067	오산시티자이1.2단지	223000336.0

3. 격자(gid)를 기준으로 groupby_sum하여 중복되는 격자를 정리

내부 데이터 전처리 - Point type & MultiPolygon

오산시_유동인구

1. 특정 지역의 월평균 유동인구 산출 -> 모든 시간대의 유동인구 데이터 합
2. 위,경도(lat, lon)를 기준으로 연평균 데이터로 변환
※ 각 지점마다 결측월이 있는 경우는 평균값으로 대치
3. 위,경도의 좌표값을 기준으로 '오산시_어린이교통사고' 격자에 **join**

차량등록현황, 총거주인구, 유소년거주인구

1. '오산시_어린이교통사고_격자' 데이터셋과 격자를 기준으로 하여 **join**함

내부 데이터 전처리 - Line String type

추정교통량, 혼잡빈도강도, 혼잡시간강도

1. 전일 시간대 데이터 선택

2. 상행과 하행을 하나로 인식 (link_id 의 마지막2자리 제외 : 01 - 상행, 02 - 하행)

3. 각 link_id가 가지는 속성을 파악하기 위해 '23.오산시_상세도로망_Level6.csv' 상세도로망 link_id로 병합

※ 고속도로, 도시고속국도, 고속도로 연결램프는 분석목적(어린이 교통사고)와 맞지 않는다고 판단해 제외 (코드정의서 참고, road_rank 에서 해당하는 숫자 제외)

4. 도로 데이터는 LineString이기 때문에 도로폭을 고려하여 버퍼적용

5. 격자의 중심점을 지나는 도로의 데이터를 측정

※ 한 격자에 두개 이상의 도로가 있는 경우 도로의 교통량, 강도를 합하거나 평균을 냄

6. '오산시_어린이교통사고_격자' 데이터셋과 격자를 기준으로 하여 **join**함

내부 데이터 전처리 - Point type & MultiPolygon

<파생변수> : 주정차 단속+횡단보도

: 횡단보도 근처의 불법 주정차 차량이 보행자와 운전자의 시야를 가려 사고를 유발

1. 횡단보도 면적에 버퍼(0.00005, 5m) 적용

2. 버퍼가 적용된 횡단보도의 면적에 포함된 주정차 단속 좌표의 개수를 count

3. 격자별로 groupby_sum

4. '오산시_어린이교통사고_격자' 데이터셋과 격자를 기준으로 하여 **join**함

내부 데이터 전처리 - Point type

<파생변수 : Region_type(거주지역) >

-
1. '오산시_도로명주소_건물' 데이터를 주거용, 사무용, 기타로 다변주화

Ex) region_type = HS / OF /OTH

2. 다변주화된 region_type을 one-hot encoding을 통해 새로운 파생변수 생성

Ex) region_type_HS, region_type_OF, region_type_OTH

<파생변수 : float_type(유동인구) >

-
1. '오산시_유동인구' 데이터를 quantile을 고려하여 유동성이 높은 지역, 중간 지역, 낮은 지역

Ex) float_type = H / M / L

2. 다변주화된 region_type을 one-hot encoding을 통해 새로운 파생변수 생성

Ex) float_type_H, float_type_M, float_type_L

내부 데이터 전처리 - **어린이_교통사고**

: 종속변수로써 사용

데이터 불균형

원데이터의 사고건수는 0인 row의 비율이 매우 높은 불균형한 분포를 가지고 있음

**공간자기상관성**

공간데이터의 특성상 한 지역에서 일어난 사건이 그 주변 지역에까지 영향을 미침

불균형한 데이터를 그대로 예측하게 된다면
과적합이 발생할 가능성이 높아짐

“원본 데이터”**“교통사고_가중치”**

“거리별로 버퍼를 활용한 가중치 적용”

내부 데이터 전처리 - 어린이_교통사고

: 종속변수로써 사용

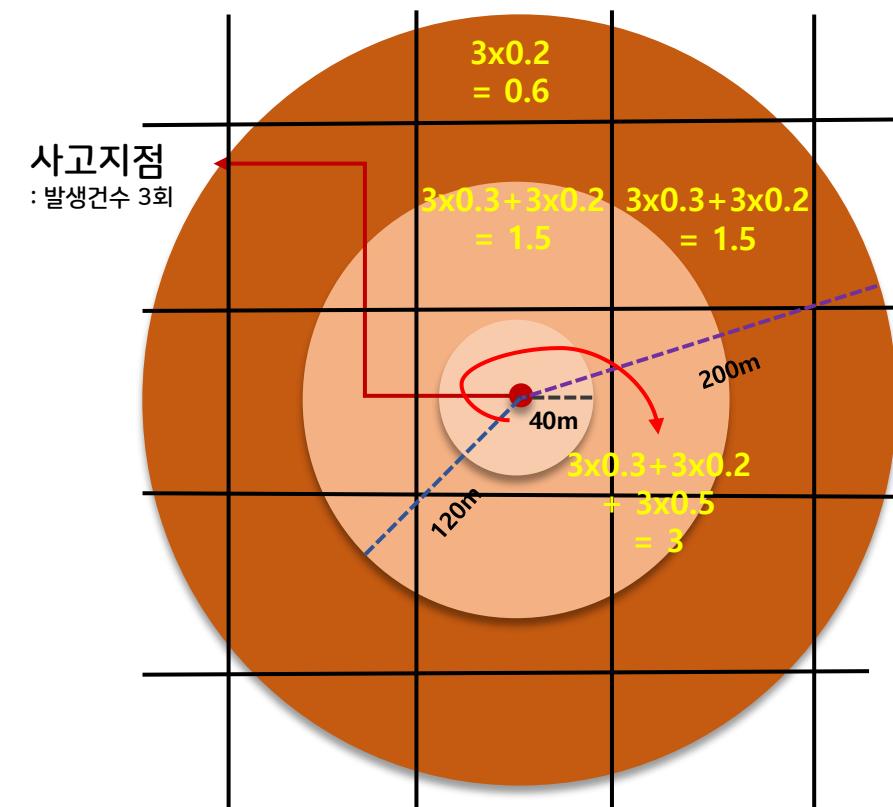
“원본 데이터”



“교통사고_가중치”

“거리별로 버퍼를 활용한 가중치 적용”

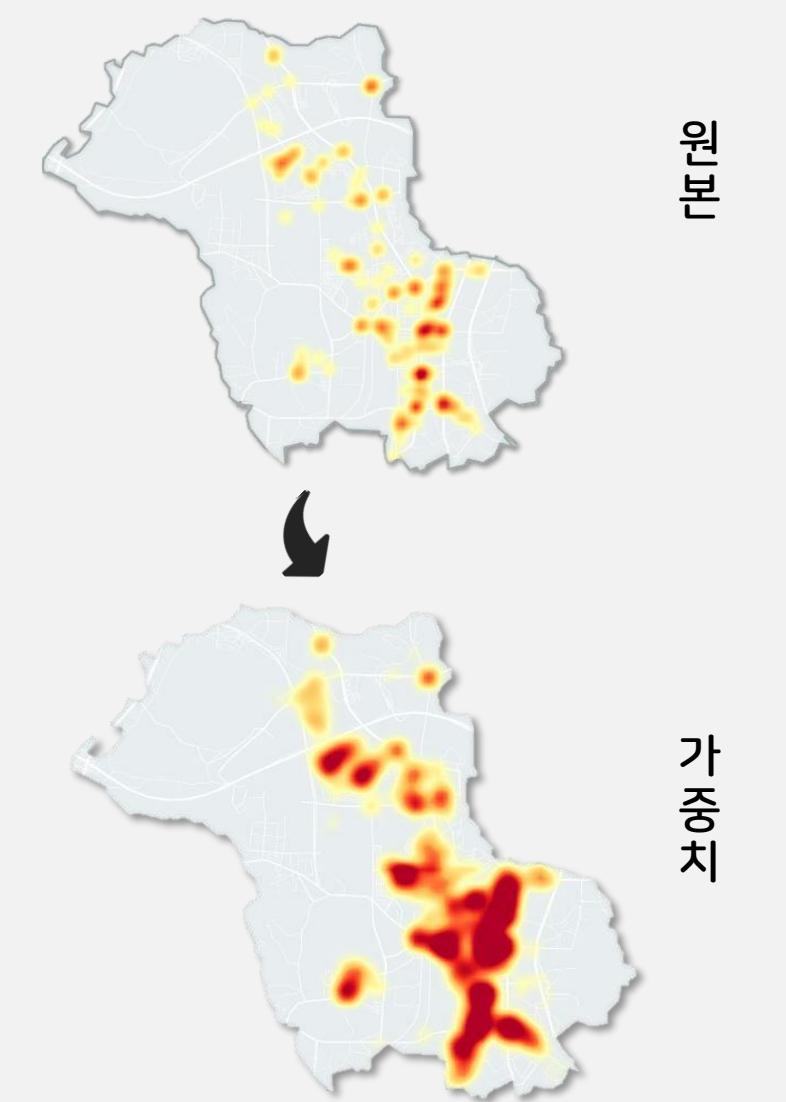
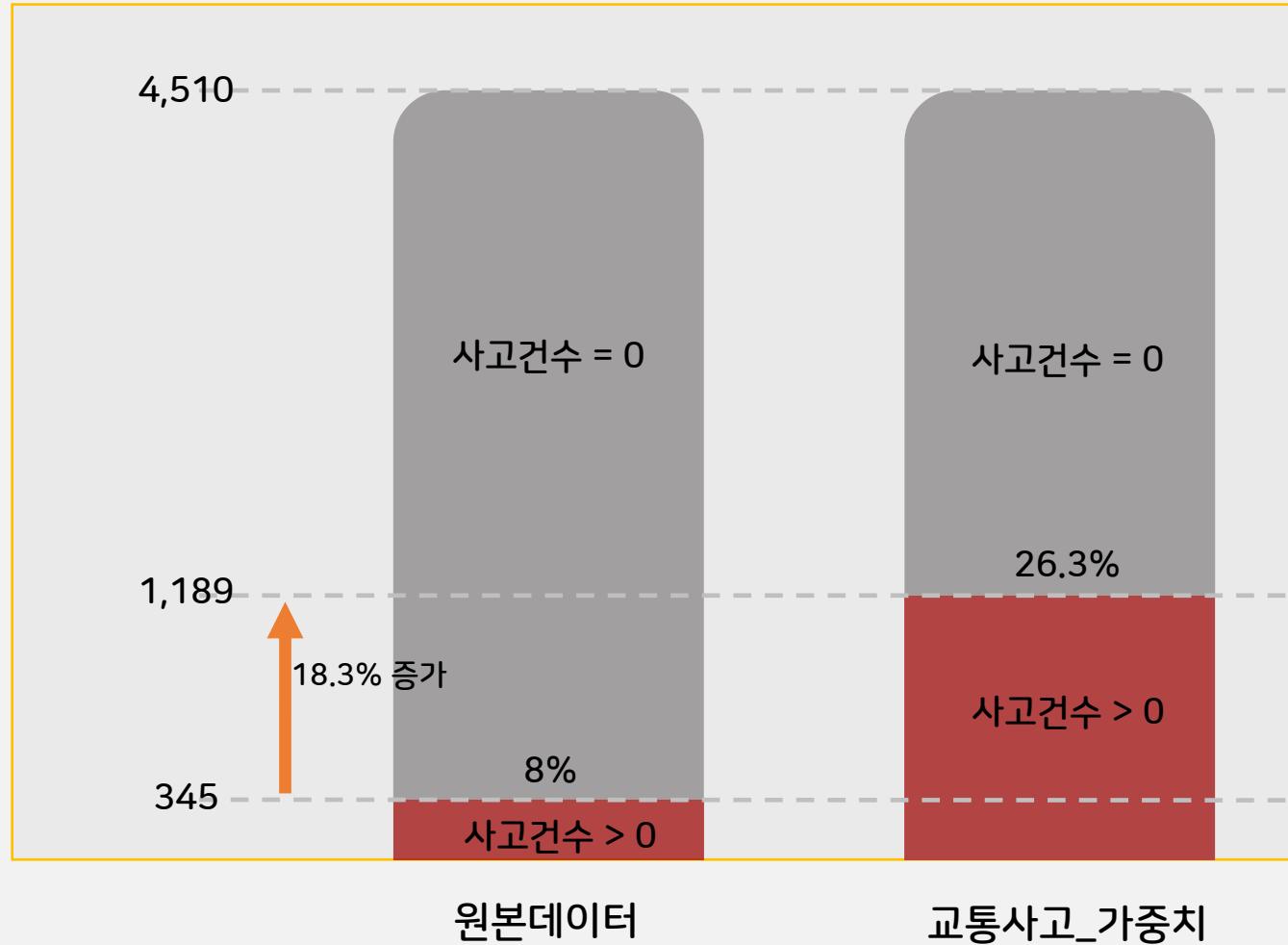
- 사고지역으로부터 역거리가중법을 부여한
새로운 종속변수를 생성
- 기존 정보의 유실을 최대한 막고, 불균형을
해결하고, 공간의 영향까지 고려



내부 데이터 전처리 - 어린이_교통사고

: 종속변수로써 사용

원본데이터와 가중치적용 데이터의 불균형 완화



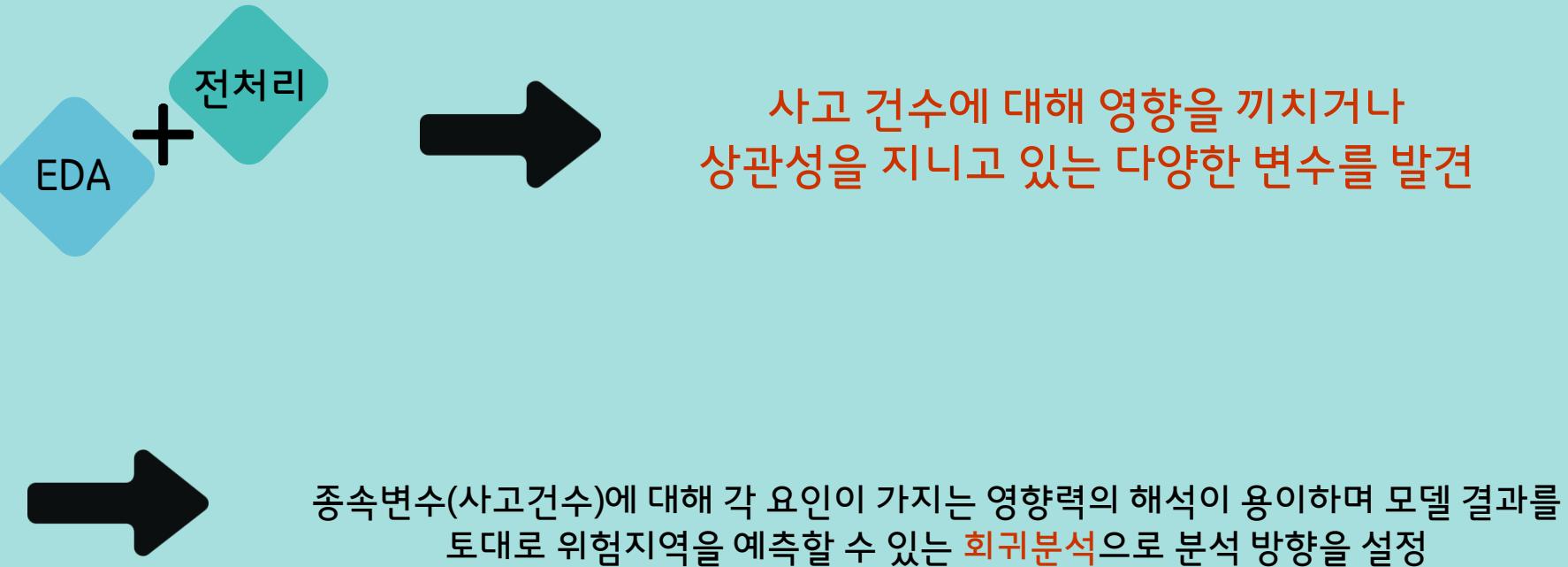
모델링



분석 방향

Problem

어린이 교통사고에 영향을 끼치는 요인을 파악하고,
높은 사고 발생건수가 예측되는 위험지역을 도출



회귀분석

OLS



공간회귀분석(Spatial Regression)
: 각 요인의 영향을 해석하고 변수를 선택

SEM

SLM



사고의 예측력을 높이기 위해
다양한 머신 러닝과 앙상블을 활용하여 예측모델 개발

knn

svm

LGBM

...

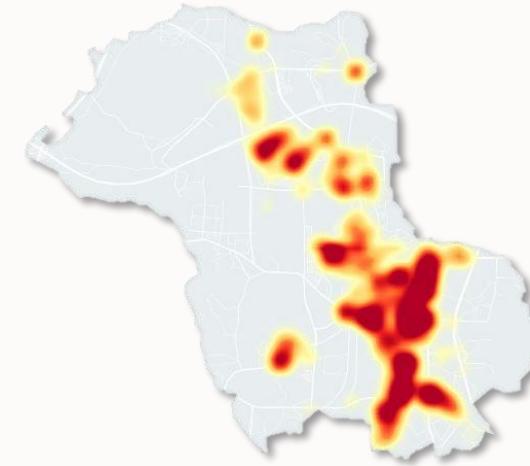
Random
Forest



공간회귀분석



어린이 교통사고와 같이 공간상의 한 위치에서 발생하는 사건은 그 주변지역에서 발생하는 사건과 높은 상관관계를 가짐(공간자기상관성, Spatial autocorrelation)



따라서 단순한 OLS모형으로는 공간적 특성을 지닌 데이터를 설명하기에 충분하지 못하며
왜곡된 추정결과를 낳으므로 이를 통제하기 위해 공간회귀모형을 사용

공간시차모형
(spatial lag model; SLM)

공간오차모형
(spatial error model; SEM)

“

위와 같은 공간회귀모형을 사용해 OLS 모형의 한계를 극복하고,
이를 토대로 독립변수가 어린이 교통사고에 미치는 영향을 분석하기로 함

”

Step 1. 다중공선성 확인

단순 OLS 모형을 이용해
변수들 간의 다중공선성을 확인

	VIF Factor	features
0	3.896033	const
1	1.269417	단속카메라
2	1.646612	도로안내표지
3	1.216494	과속방지턱
4	1.551880	스쿨존
5	2.527156	횡단보도_전체면적
6	2.290877	신호등
7	1.513835	단속횟수_횡단보도위
8	1.749845	단속횟수
9	2.280357	traffic_sum
10	41.675152	freq_sum
11	50.219174	time_sum
12	1.254150	초등학교_100m
13	2.381285	유치원_100m
14	1.983553	유동인구
15	1.088932	cctv
16	1.177781	버스정류장
17	1.266076	체육시설
18	1.258905	학원
19	1.160591	car_cnt
20	11.853514	pop_all



'time_sum = 혼잡시간강도' ,
'pop_all = 거주인구격자(총인구)'
변수 제거

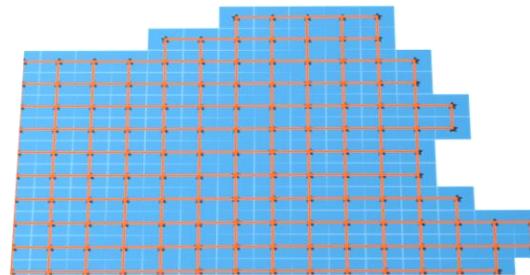
	VIF Factor	features
0	3.896576	const
1	1.271691	단속카메라
2	1.635744	도로안내표지
3	1.188827	과속방지턱
4	1.737141	스쿨존
5	2.434051	횡단보도_전체면적
6	2.182269	신호등
7	1.512172	단속횟수_횡단보도위
8	1.739924	단속횟수
9	2.050497	traffic_sum
10	2.085377	freq_sum
11	1.900295	초등학교_가중치
12	2.267899	유치원_가중치
13	1.955626	유동인구
14	1.074482	cctv
15	1.174524	버스정류장
16	1.260068	체육시설
17	1.260380	학원
18	1.145675	car_cnt
19	1.846257	pop_child
20	1.476843	region_type_HS

Step 2. 공간자기상관성 확인 - 공간가중행렬

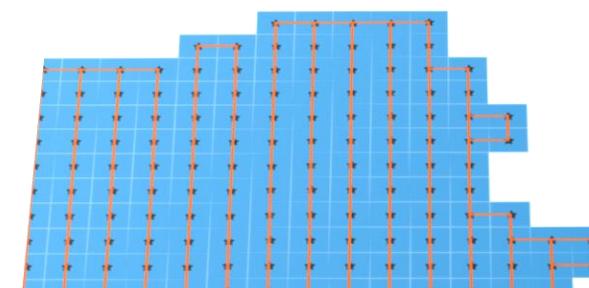
공간자기상관성을 확인하기 위해 **공간가중행렬**을 정의해야함

1) 인접성척도를 기준으로 정의하는 방식

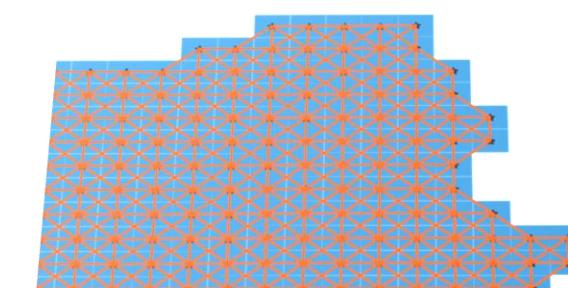
Rook



knn



Queen



2) 거리척도를 기준으로 정의하는 방식

- 500m를 기준으로 거리의 역수로 가중치를 부여한 행렬 정의

“두 가지 방식 모두 행의 합이 1이 되도록 정규화”

Step 2. 공간자기상관성 확인 – Moran's I

Moran's I : 공간자기상관성을 확인하는 지표

DIAGNOSTICS FOR SPATIAL DEPENDENCE			
TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.0664	8.924	0.0000
Lagrange Multiplier (lag)	1	45.749	0.0000
Robust LM (lag)	1	2.425	0.1194
Lagrange Multiplier (error)	1	76.070	0.0000
Robust LM (error)	1	32.746	0.0000
Lagrange Multiplier (SARMA)	2	78.495	0.0000

p-value가 매우 작고 Z값이 8에 가까운 값이므로 공간자기상관성이 있다고 할 수 있음

공간회귀모형을 적용하는 것이 적절함

Step 3. 모형선택 – Robust LM

Robust LM : SLM과 SEM 모형중 어떤 것이 더 적절한지 판단하는 지표

DIAGNOSTICS FOR SPATIAL DEPENDENCE			
TEST	M1/DF	VALUE	PROB
Moran's I (error)	0.0389	14.277	0.0000
Lagrange Multiplier (lag)	1	114.599	0.0000
Robust LM (lag)	1	5.042	0.0247
Lagrange Multiplier (error)	1	186.823	0.0000
Robust LM (error)	1	77.265	0.0000
Lagrange Multiplier (SARMA)	2	191.864	0.0000

Lag(SLM)과 error(SEM)의 지표 모두 유의하지만, error의 유의확률의 값이 더 낮음

공간오차모형(SEM)을 선택

Step 4. 회귀계수 추정 방식 선택

회귀계수를 추정하는 방식은 ML방식과 GM방식으로 나뉨



: 오차의 정규성을 만족하지 못함



: 정규성 가정 필요 x

GM방식 선택

Step 5. 변수조합

1. 공간오차모형(SEM) + GM + 공간가중행렬(Queen)
2. 공간오차모형(SEM) + GM + 공간가중행렬(거리척도)



두가지 공간가중행렬 방식을 각각 사용하고,
총 16가지의 변수 조합에 대한 분석을 시도함

→ 공간회귀분석 32회 실시

Step 6. 최종변수선택

SEM_GM_Queen		R2	유의하지 않은 변수	
초등학교_100m	유치원_100m	0.3519	과속방지턱, 초등학교, 유치원, 유동인구, pop_child, HS, OF, H, M	
	유치원_300m	0.3550	과속방지턱, 초등학교, 유동인구, pop_child, HS, OF, H	
	유치원_500m	0.3588	과속방지턱, 스쿨존, 초등학교, 유동인구, pop_child, HS, OF, H	
	유치원_가중치	0.3519	과속방지턱, 초등학교, 유치원, 유동인구, pop_child, HS, OF, H, M	
.				
.				
.				
SEM_GM_DistanceBand		R2	유의하지 않은 변수	
초등학교_300m	유치원_500m	0.3616	과속방지턱, 스쿨존, 유동인구, pop_child, HS, OF	
	.			
	.			
	.			
초등학교_가중치	유치원_100m	0.3576	과속방지턱, 유치원, 유동인구, car_cnt, pop_child, HS, OF, H	
	유치원_300m	0.3585	과속방지턱, 스쿨존, 유동인구, pop_child, HS, OF, H	
	유치원_500m	0.3610	과속방지턱, 스쿨존, 유동인구, pop_child, HS, OF,	
	유치원_가중치	0.3576	과속방지턱, 유치원, 유동인구, car_cnt, pop_child, HS, OF, H	

32개의 모델중에 가장 설명력이 높은 모델을 선택

→ “SEM_GM_DistanceBand_초등학교_300m_유치원_500m”

선택된 모델에서 유의한 회귀계수를 가진 변수를 선택함

변수 list

'단속카메라'
 '도로안내표지'
 '횡단보도_전체면적'
 '신호등'
 '단속횟수'
 '단속횟수_횡단보도위',
 'traffic_sum'
 'freq_sum'
 '초등학교_300m'
 '유치원_500m'
 'cctv','버스정류장'
 '체육시설'
 '학원'
 'car_cnt'
 'float_type_H'
 'float_type_M'





합성데이터



공간회귀분석을 통해 얻어진 변수들로 본격적인 머신 러닝을
진행하기에 앞서 모델을 학습시킬 데이터를 확보 해야함

**하지만 전처리를 통해 얻은 원본 데이터는 위험지역을 도출하는데
사용되는 위험지역 예측 데이터이기 때문에 동일한 데이터로 모델을 학습시킬 수 없음**



따라서 원본 데이터의 특성 분포를 따르면서
기존 데이터와 일치하지 않는 새로운 학습용 데이터를 확보하기 위해
“SMOTE” 기법 사용

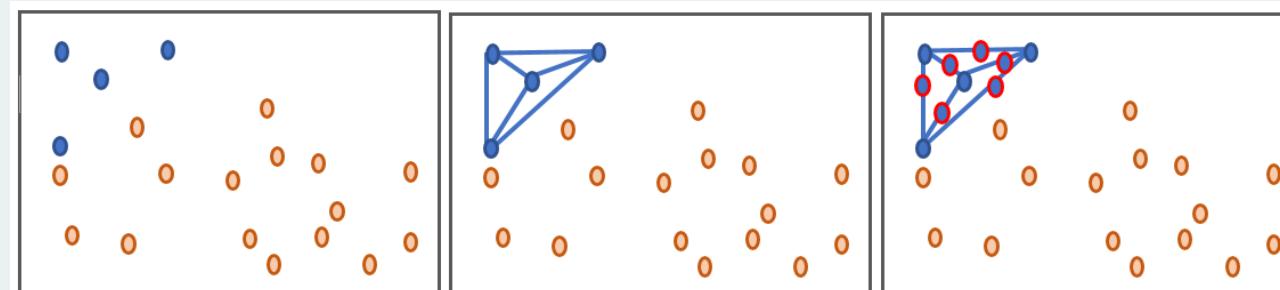
SMOTE : Synthetic Minority Over-sampling Technique

SMOTE는 데이터 불균형을 해소하기 위해 고안된 오버샘플링 기술이다.

합성 소수 샘플링 기술을 사용해 다수 클래스를 샘플링하고 기존 소수 샘플을 보간하여 새로운 소수 인스턴스를 합성해낸다.

동작방식

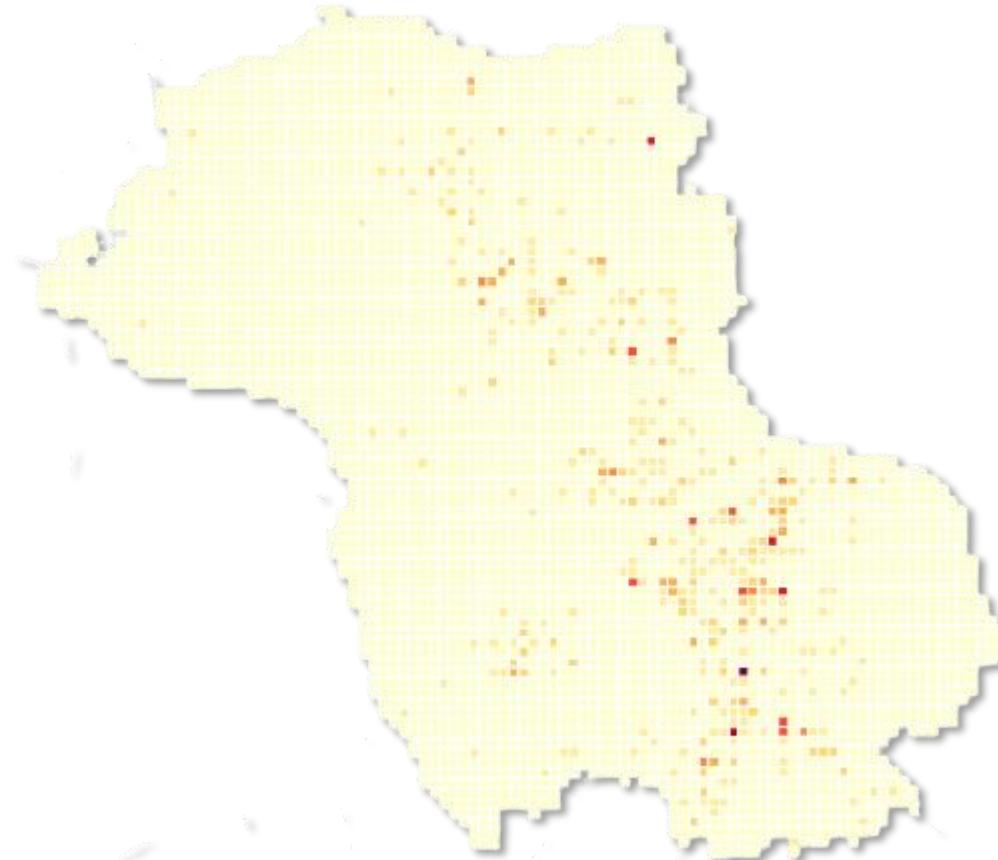
1. K-nearest-neighbor method를 활용하여 샘플 포인트와 근접한 k개의 샘플을 선정
2. 소수 데이터 중 특정 샘플과 가장 가까운 이웃 사이의 차이(difference)를 계산
3. 이 차이에 0과 1사이의 난수를 곱함
4. 원본 데이터에 difference와 난수를 곱한 값을 더하여 새로운 합성데이터를 생성함



출처 - [@DataSciCampus](#)

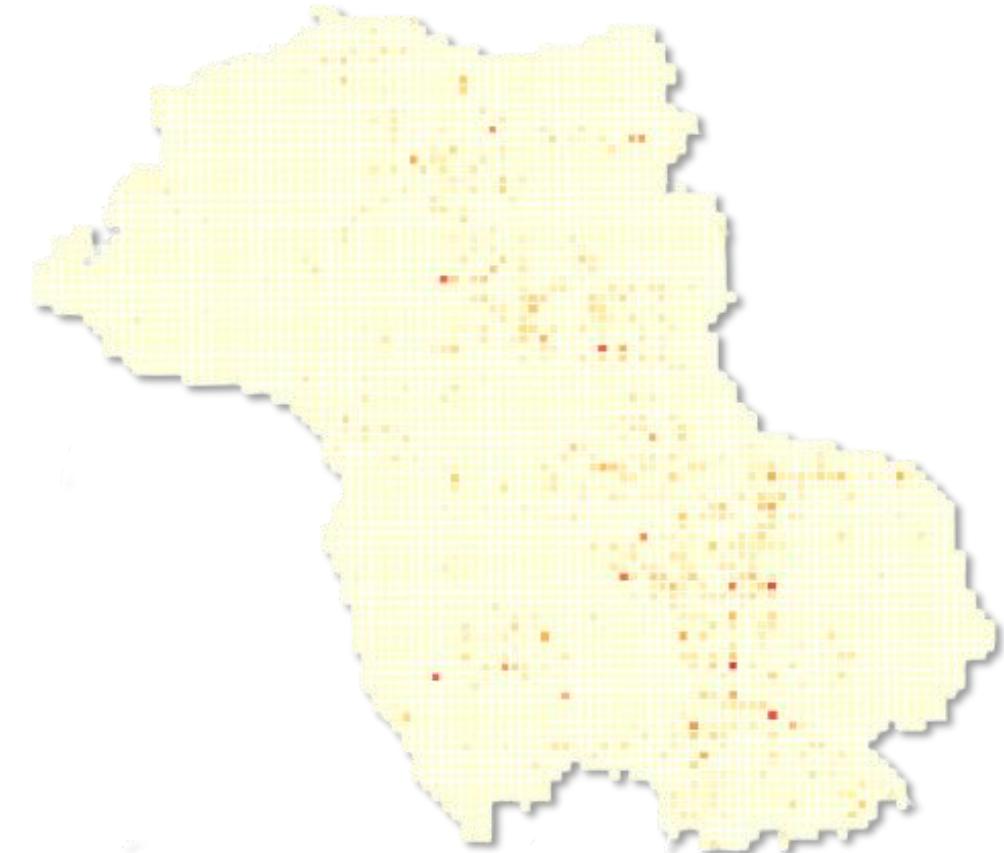
원본데이터

사고횟수_가중치



합성데이터

사고횟수_가중치_합성

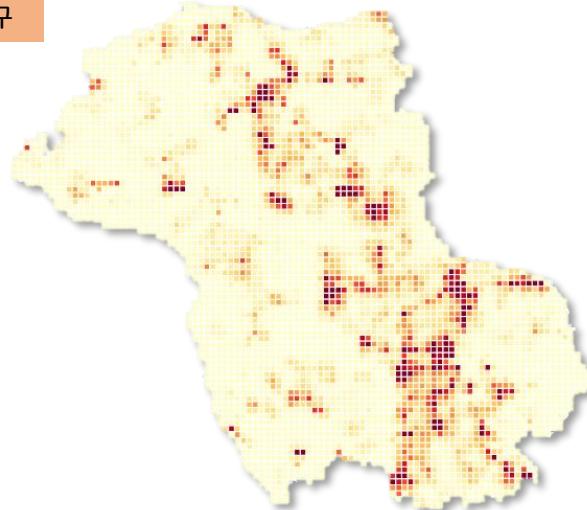


모델링

: Synthetic Data (합성데이터) - (4)

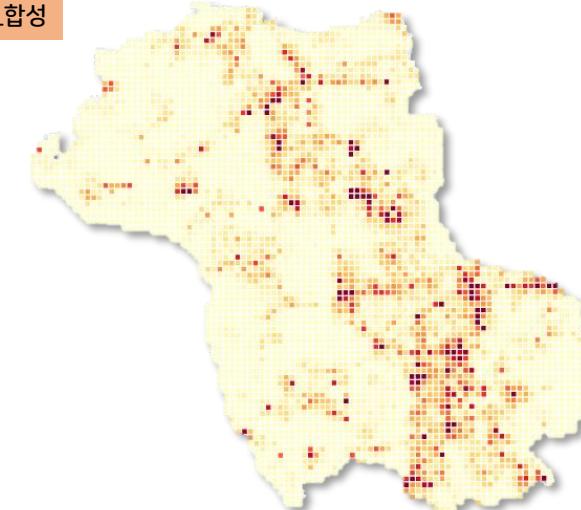
원본데이터

유동인구

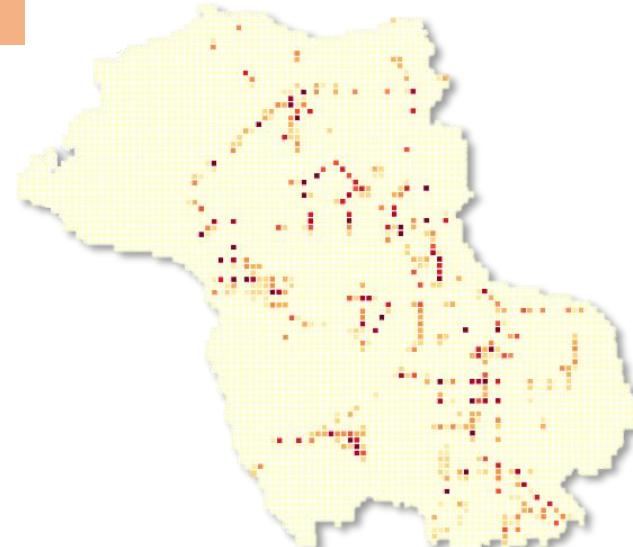


합성데이터

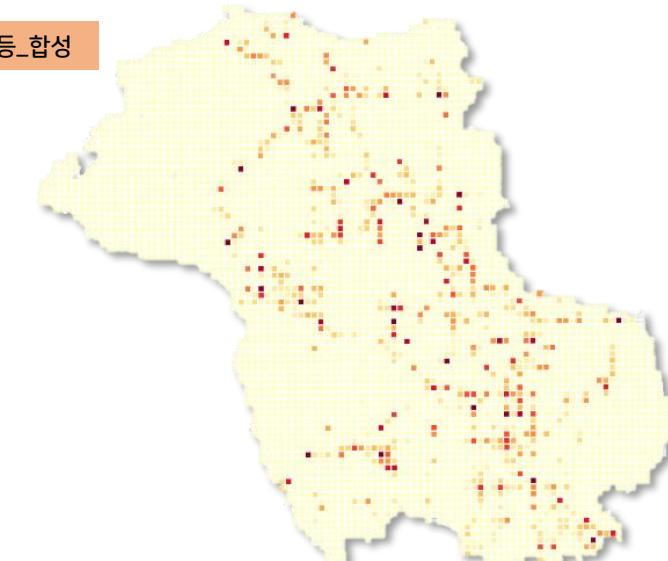
유동인구_합성



신호등



신호등_합성

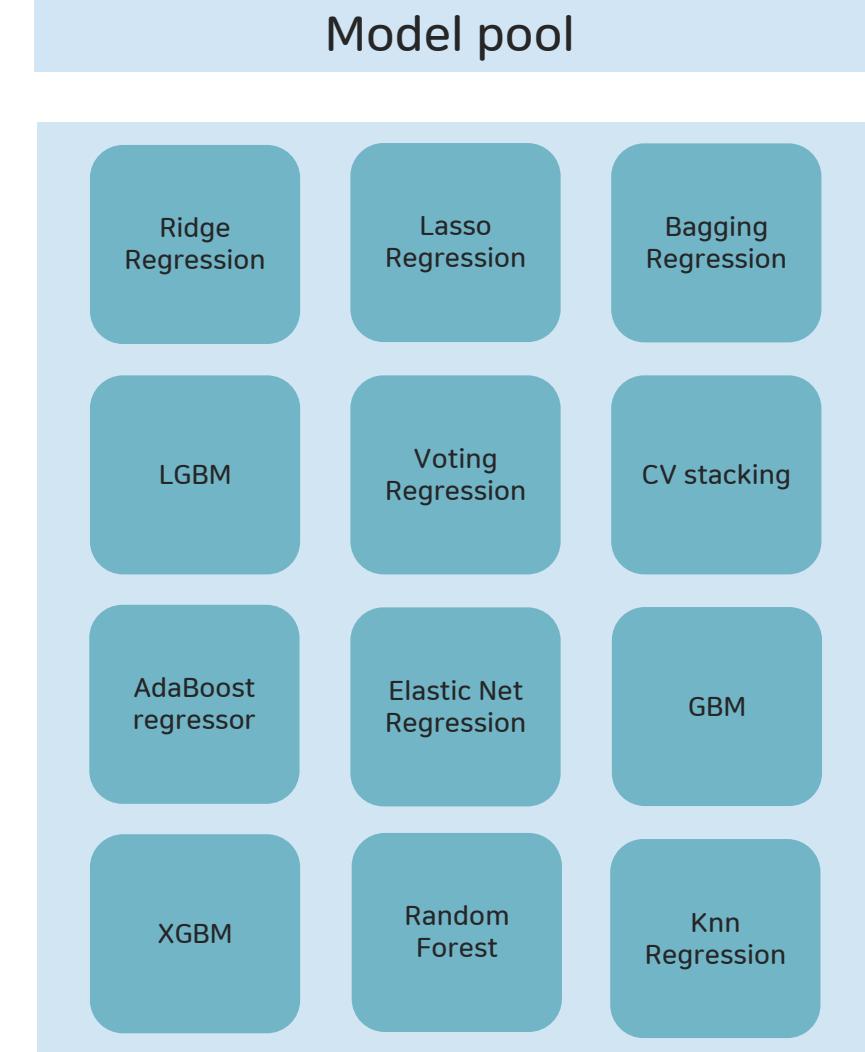
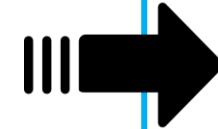
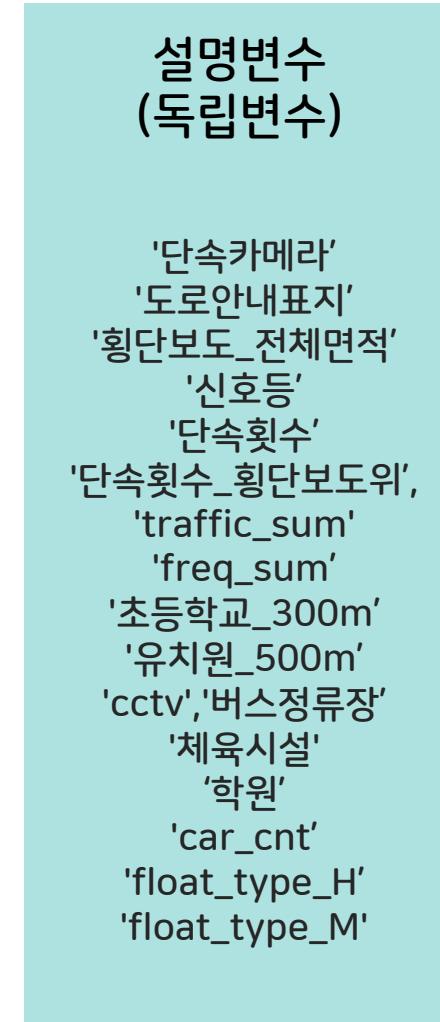
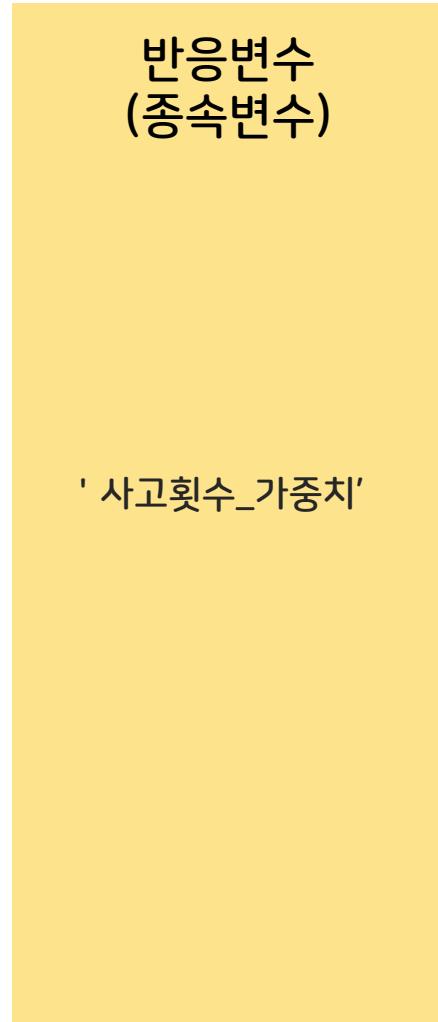




머신러닝



최종적으로 생성된 학습 데이터셋으로 다양한 모델들을 학습시킴



“1차적으로 모형의 성능을 확인하여 모델을 선별”

Algorithm	Train R^2	Test R^2	Test Rmse
Lasso regression	0.544	0.573	0.928
Gradient boosting regression	0.804	0.718	0.754
Ridge regression	0.613	0.638	0.855
Knn regression	0.646	0.410	1.092
RF regression	0.953	0.742	0.721
Adaboost + Decision tree	0.985	0.731	0.737
Bagging + decision tree	0.932	0.729	0.739
Elastic net regression	0.553	0.588	0.912
LGBM	0.915	0.740	0.725
XGBM	0.973	0.741	0.723

선별된 모델



- 의사결정나무 기반의 앙상블(Bagging) 모형
- 모형이 간단하고 결과 해석이 용이함
- overfitting에 취약하지 않음



- 결정 트리 기반의 앙상블(Boosting) 모형
- 리프 중심 트리분할 방식 사용 / 뛰어난 예측 성능 / 개선된 연산 시간



- 결정 트리 기반의 앙상블(Boosting) 모형
- 균형 트리분할 방식 사용 / 뛰어난 예측 성능

“모델의 예측력을 높이기 위해 파라미터 튜닝 진행”

algorithm	Best_Parameter					
RF regression	n_estimators = 400		max_depth = 10	min_samples_split = 100		
LGBM	Learning_rate = 0.05		max_depth = 10		num_iterations = 500	Num_leaves = 8
XGB	base_score = 0.5	booster = 'gbtree'	colsample_bylevel = 1	colsample_bynode = 1	n_jobs = 0	num_parallel_tree = 1
	reg_alpha = 0	reg_lambda = 1	scale_pos_weight = 1	subsample = 1	verbosity = None	tree_method = 'exact'

*random_state=123

“튜닝이 완료된 모델들을 대상으로 Ensemble 기법 사용”

Ensemble

: 여러 가지 우수한 학습 모델을 조합해 예측력을 향상시키는 모델

- 장점 - 단일 모델에 비해 예측 성능 우수
- 학습 데이터가 적을 때 모델을 학습하기 용이함

- 단점 - 예측 시간이 오래 걸림
- 결과의 해석이 어려움



- 분석 과제는 어린이 교통 사고 위험지역 예측이 최우선 과제
- 전체 데이터셋의 관측치는 약 4510개로 학습 데이터가 적음
- 공간회귀분석을 통해 모델의 예측결과 해석 보완

따라서 다양한 Ensemble 모형으로 예측을 진행

원본 데이터 셋을 테스트 데이터셋으로 하여 최종 앙상블 모델 결과 비교

	합성데이터의 Test Set		전체 원본데이터(4510 obs)	
	RMSE	MAE	RMSE	MAE
Random Forest	0.7187	0.4338	0.9178	0.5095
LGBM	0.7357	0.4433	0.9116	0.5118
XGB	0.7145	0.4326	0.9205	0.5091
Voting Reg	0.7099	0.4304	0.9035	0.5042
Stacking	0.7681	0.4571	0.9449	0.5245

“
최종 모형 Voting Regressor 선택
 (Voting Regressor는 RandomForest, LGBM, XGB 세 모형을 Soft Voting한 모형)
 ”

*soft voting = 여러가지 개별 모형의 평균 예측값을 사용하는 방식



모델링

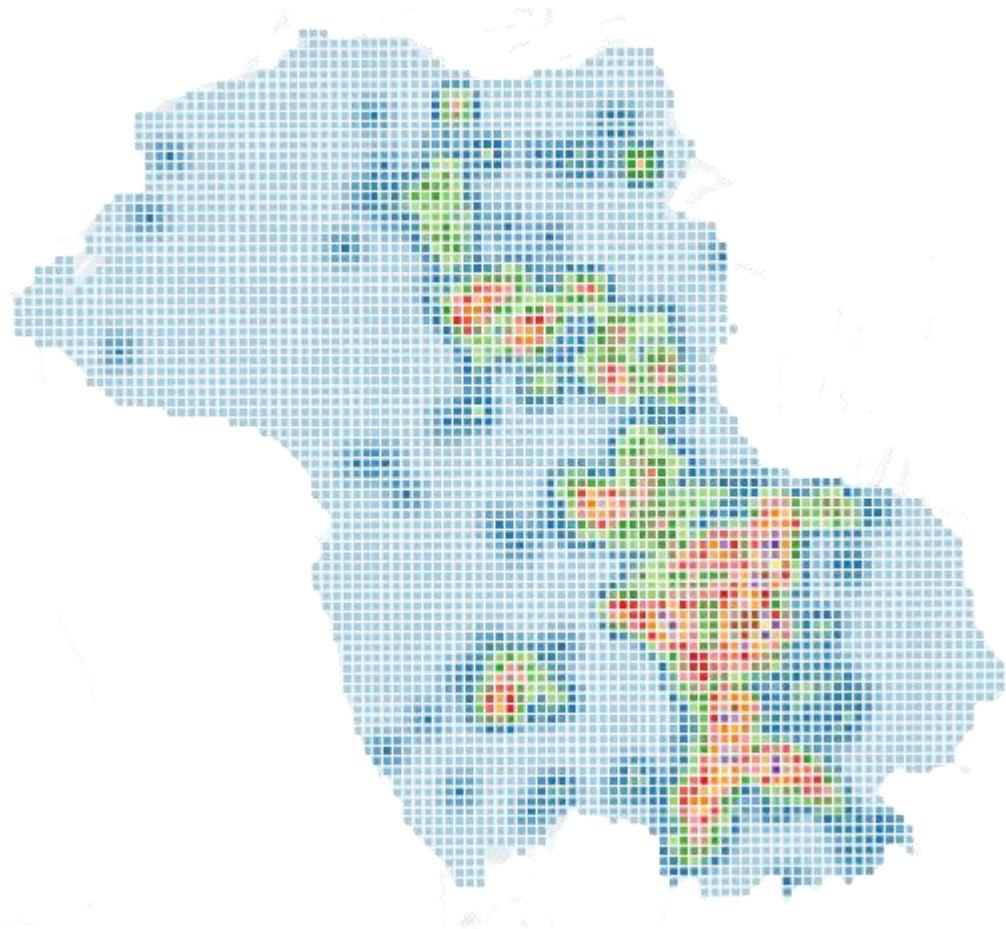
: 머신러닝 - (7)

최종모델예측결과

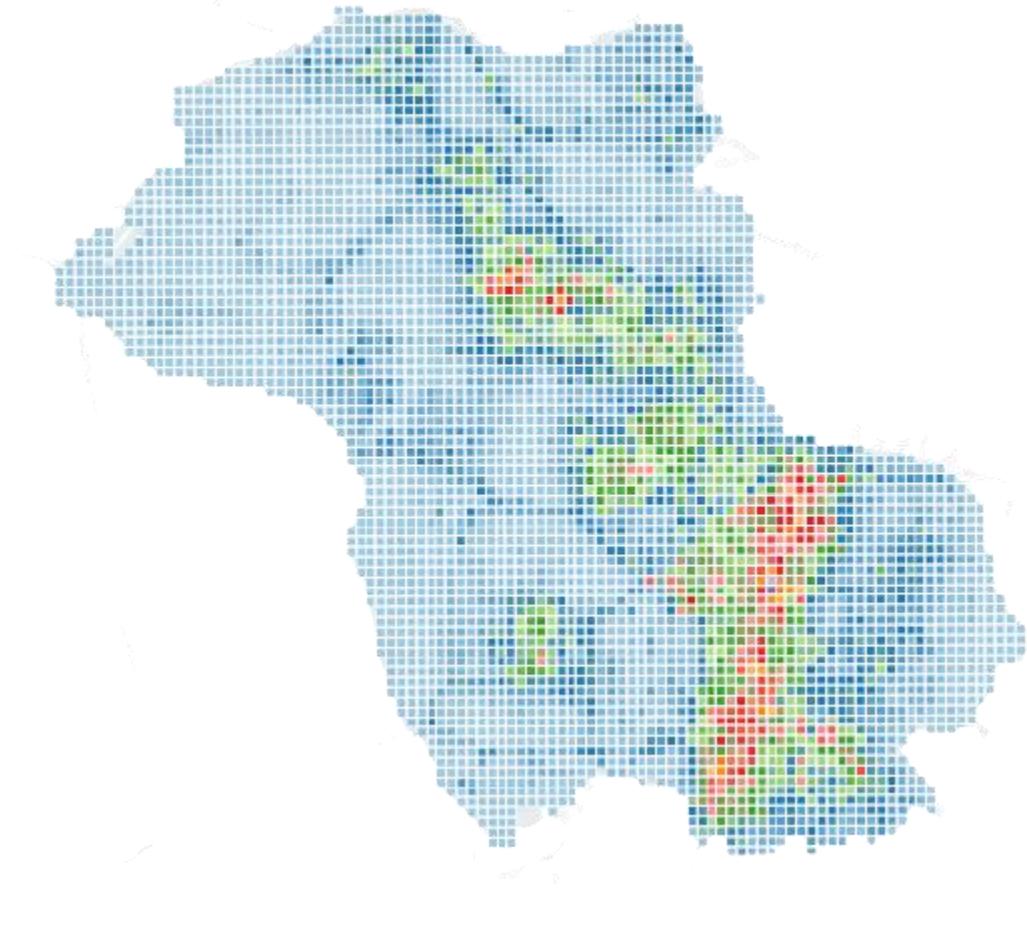
: predicted 값을 내림차순으로 정렬

사고 횟수 가 중치	predicted	단 속 카 메 라	도 로 안 내 표 지	횡단보도_전 체면적	신 호 등	단속횟수	단속 횟수 평 균 단 보 도 위	traffic_sum	freq_sum	초등학 교 _300m	유치원 _500m	cctv	버 스 정 류 장	체 육 시 설	학 원	car_cnt	float_type_H	float_type_M	
gid																			
다사 622056	13.3	8.683458	1	3	2.386043e-08	7	392.0	369.0	54555.53	456.09	0	27	0	0	2	7	19	1	0
다사 622057	11.8	7.999273	1	0	1.860388e-08	0	1071.0	1.0	27175.25	504.86	0	29	0	0	0	10	23	1	0
다사 625064	9.0	7.602250	1	4	3.605260e-08	8	267.0	534.0	37142.47	550.65	1	37	1	0	3	4	12	1	0
다사 623057	11.4	7.469966	0	0	1.862217e-08	5	465.0	400.0	32057.33	712.62	1	29	0	0	2	3	6	1	0
다사 620049	7.4	7.307602	1	1	1.785115e-08	3	50.0	2.0	32716.22	489.54	2	27	0	0	0	0	17	1	0
다사 622055	11.3	7.262991	0	0	1.853523e-08	0	70.0	1.0	12527.53	495.11	0	24	0	0	0	0	3	1	0
다사 623056	11.7	7.100805	0	4	3.200839e-08	8	19.0	0.0	57483.47	472.54	0	27	0	0	0	2	35	1	0
다사 617039	7.1	7.039752	0	0	1.400709e-08	8	3.0	0.0	36585.05	515.60	2	36	0	1	0	0	7	1	0
다사 624062	9.5	7.038032	0	0	9.355671e-09	4	0.0	0.0	19429.70	148.17	2	36	0	1	0	0	3	1	0
다사 621044	7.8	6.895137	1	0	0.000000e+00	0	20.0	0.0	54197.66	587.30	1	29	0	0	0	1	112	1	0
다사 617038	6.1	6.739863	0	0	6.094368e-09	0	72.0	5.0	30626.11	385.15	2	37	0	1	1	1	13	1	0
다사 624056	9.9	6.725058	0	1	2.289070e-08	0	0.0	0.0	43171.20	378.32	0	26	0	1	0	0	0	1	0

최종모델예측결과 - 시각화



Y_label
(사고횟수_가중치_원본)



Predicted value

결론



해결과제 1. 위험지역 20개소 선정

결론 - 1

: 위험지역 선정 - (1)

최종 모델의 사고 발생 건수 예측 값을 내림차순으로 정렬



기존에 어린이 보호구역으로 지정된 지역은 제외 해야함

But, 어린이 보호구역 데이터는 주변 유치원, 초등학교의 중심좌표로 설정되어 있음

정확한 어린이 보호구역에 해당하는 위치 값을 알 수 없음



예측된 위험지역 상위 130개 지역에 대해

도로, 시설물을 직접 확인해 어린이 보호구역 설정여부를 확인함

어린이 보호구역 여부
column 추가

보호구역여부
0
0
1
0
0
.
.
.
1
1
0

결론 - 1

: 위험지역 선정 - (2)

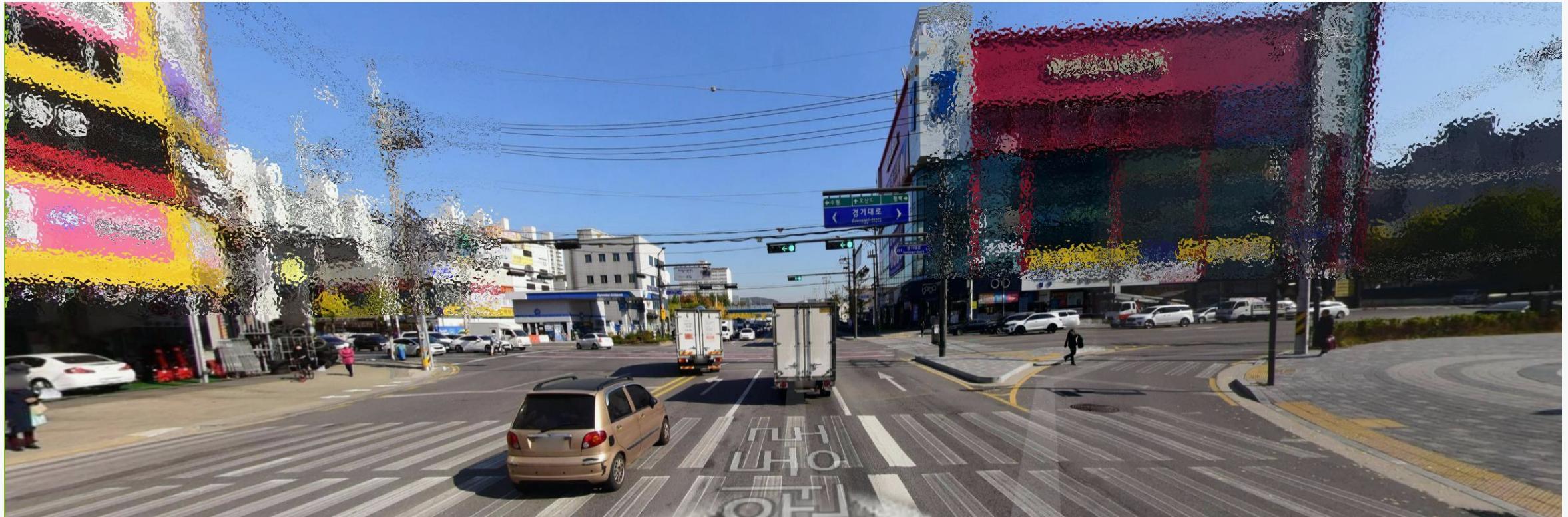
최종 어린이 교통사고 위험지역 20개소

위험순위	시설명/주소지	중심점 위치		반경범위
		X좌표(위도)	Y좌표(경도)	
1	운암뜰 한식거리 일대(성호대로)	37.148745308158986	127.07543576946158	100M
2	운암주공 4단지아파트 후문 일대	37.149885304889	127.07530385008077	100M
3	타이어뱅크 오산점 일대	37.142613988795574	127.07320839337851	80M
4	운암뜰 한식거리 내부도로	37.14754439242446	127.07532110502105	100M
5	오산농협 원동지점 일대	37.13795576847067	127.0734786288245	50M
6	투썸플레이스 오산갈곶점 일대	37.13244931931593	127.0689207864787	50M
7	오산시청 앞 교차로	37.148725911900115	127.07799836625681	80M
8	롯데마트 일대	37.14990880268169	127.07358760799065	50M
9	휴먼시아 12단지아파트 앞 상가 일대	37.17523612103733	127.05076740428677	100M
10	이마트 앞 삼거리	37.141431759443115	127.07350959079413	50M
11	운암주공 4단지아파트 입구 일대	37.15088790076599	127.07473845339443	50M
12	운암청구아파트 정문 일대(운암로)	37.14751134460182	127.07621434310592	50M
13	우주할인마트 앞 사거리	37.13790124398053	127.07503375895517	50M
14	KCC스위첸 아파트 옆 골목	37.1323226207828	127.0680978411274	50M
15	운암주공 2,3단지 사잇길	37.154777253771464	127.07768746314427	50M
16	유탑유블레스 오피스텔 앞 사거리	37.14125324161613	127.07192075662905	50M
17	MTA태권도 앞 골목(오산로62번길)	37.13175620749667	127.0694263245953	50M
18	현대오일뱅크(오산물류) 옆 삼거리	37.137916524518005	127.06879008419004	50M
19	아웃백 앞 사거리	37.146328743859186	127.0735459171848	50M
20	하나은행 오산중앙점 일대(밀머리로)	37.13935151365605	127.07469691169145	100M

결론 - 1 : 예시

: 위험지역 선정 - (3)

위험 지역 예시 - 타이어 백크



보차도가 분리되지 않음

신호등 없는 횡단보도가 존재하고 교통량이 많음

결론 - 1 : 예시

: 위험지역 선정 - (4)

위험 지역 예시 - 우주할인마트 앞 사거리



교통량이 많은 사거리 교차로에 신호등이 없음

사진 좌측 및 정면에 주정차 차량이 있어 시야 확보가 어려움

결론 - 1 : 예시

: 위험지역 선정 - (5)

위험 지역 예시 - 운암뜰 한식거리 내부도로



다수의 불법 주정차 차량이 있어 시야 확보가 어려움

코너가 많음에도 반사경이 없어 사고 위험이 높음



해결과제 2. 안전시설물 우선 설치 지역 제시

최종 모델의 사고 발생 건수 예측 값을 내림차순으로 정렬



'보호구역 여부' column을 참고해 어린이 보호구역인 지역 20개소 선정

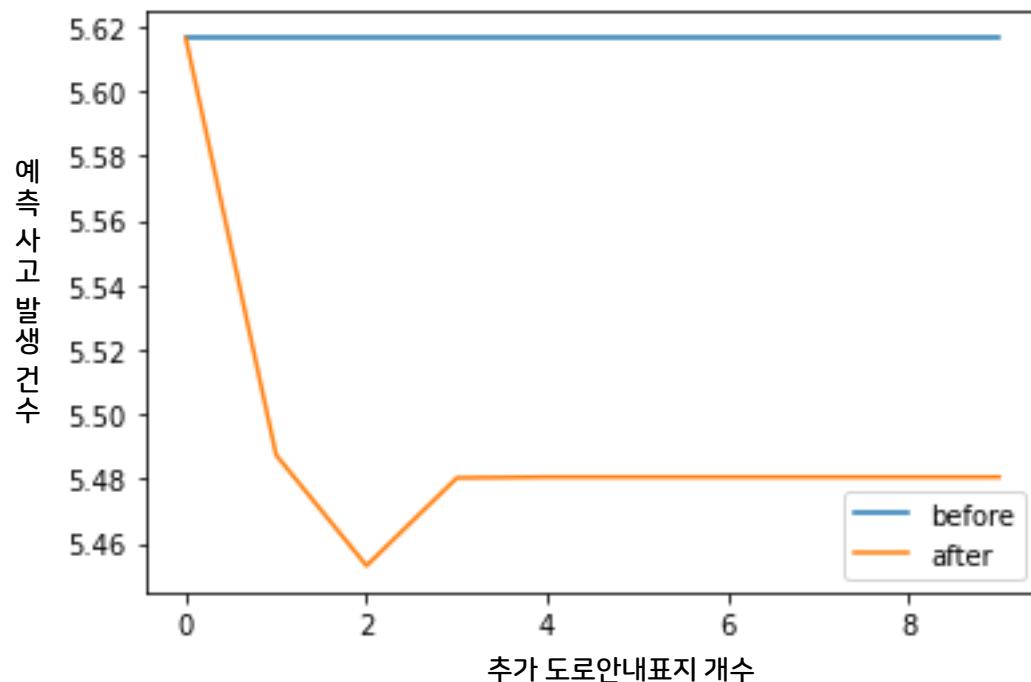


Problem : 각 지역에 대해 설치해야 할 시설물의 종류와 위치 선정

Solution : 1. 최종 모델을 활용하여 사고 건수를 감소시키는 시설물을 도출
2. 선행연구를 참고하여 도로 특성에 따른 유의미한 안전시설물 설치 제안

Solution 1. 모델활용

도로안내표지의 설치 개수 변화에 따른 어린이 교통사고 예측 발생 건수의 변화



좌측의 그래프는 예측 사고 발생 건수 상위 20개 지역에 대해
도로안내표지를 추가 설치할 때의 평균 예측 사고 발생 건수 그래프이다.

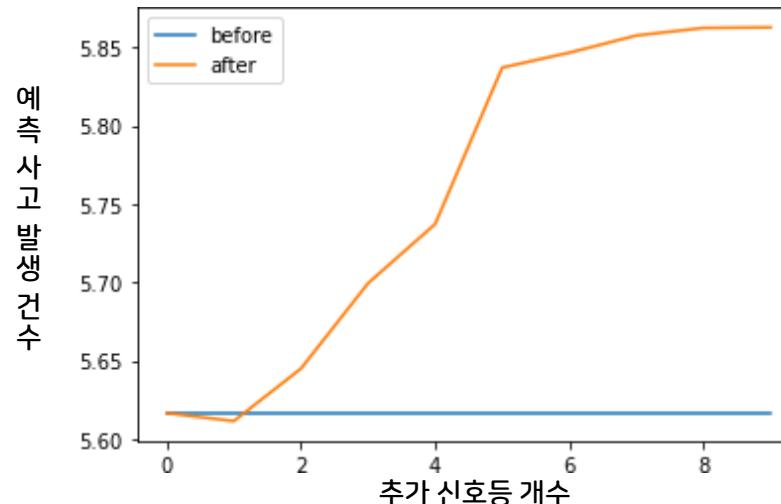
평균적으로 도로안내표지를 1~2개를 추가로 설치할 때, 사고 감소의 효과를 얻음

결론 - 2

: 안전시설물 설치 - (3)

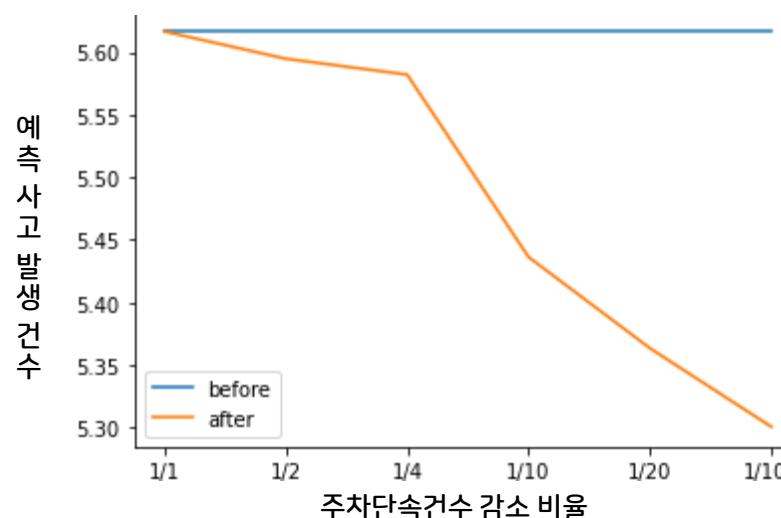
Solution 1. 모델활용

신호등의 설치 개수 변화와 단속횟수의 감소에 따른 어린이 교통사고 예측 발생 건수의 변화



좌측의 그래프는 예측 사고 발생 건수 상위 20개 지역에 대해
신호등을 추가 설치할 때의 평균 예측 사고 발생 건수 그래프이다.

신호등은 평균적으로 1개를 추가로 설치할 때 사고 감소 효과를 기대할 수 있으며
2개 이상을 설치한다면 오히려 예측 사고 발생 건수가 증가하는 추세를 보인다.



좌측의 그래프는 예측 사고 발생 건수 상위 20개 지역에 대해
주차 단속 건수가 감소될 때 평균 예측 사고 발생 건수 그래프이다.

주차단속건수가 감소함에 따라 예측 사고 발생 건수 또한 감소한다.
이를 통해 불법 주정차를 줄이는 것이 사고 예방에 효과적임을 알 수 있다.

결론 - 2

: 안전시설물 설치 - (4)

Solution 2. 선행 연구 참고

안전시설물 데이터 수집의 한계와 변수들의 복잡한 선행성으로 인해 예측 모델만으로

모든 안전 시설물 종류와 위치를 결정하는 것은 한계가 있음

따라서 모델에서 제안하지 못하는 안전시설물은 선행연구를 참고하여 결정

→ 교통사고의 원인을 크게 3가지로 나누어 각 경우에 해당하는 유의미한 안전시설물의 종류와 설치 위치를 선정

과속 감소

- 과속방지턱
- 고원식 횡단보도
- *횡단보도 투광등
- 과속 단속카메라

도로 조명

- 조명시설
- 표지병

도로 형태관련

- 도류화
- 중앙분리대
- 무단횡단 방지휀스

*횡단보도를 밝게 하면 운전자가 과속을 하는 경향을 억제해줌
- 교통사고 원인 분석 및 대응방안에 관한 연구, 정우진

최종 어린이 보호구역 교통안전시설물 우선 설치 지역 20개소

위험순위	시설명/주소지	시설물설치위치		비고
		X좌표(위도)	Y좌표(경도)	
1	운천초등학교	37.156255088675195	127.07842262005786	중앙분리대(무단횡단방지시설)
2	운천고등학교	37.155238104313774	127.0801807579631	과속단속카메라
3	원동초등학교	37.135041558445806	127.06899474333302	과속단속카메라
		37.13422317372425	127.06903411900474	도로안내표지(주차장 출입구)
4	운산초등학교	37.15276279497922	127.07669226004347	과속방지턱
5	금암초등학교	37.17762911629891	127.04372890768045	보행자신호등
6	오산세교복지타운	37.17684013411641	127.04549597571159	주정차단속카메라
7	운암제1어린이공원	37.15714166663421	127.0766462264768	주정차단속카메라
		37.156883007865126	127.07657648904808	
8	원일중학교	37.13889135469192	127.07288850633138	주정차단속카메라
9	원동두산동아아파트	37.1337714251863	127.07145997589477	인도 가드레스(맞은편추가설치)
10	운암현대아파트	37.15338655364295	127.07625252602392	중앙분리대(무단횡단방지시설)
11	오산시립햇살마루도서관	37.137990835622574	127.07202059964526	주정차단속카메라
		37.138336071252816	127.07292589086413	인도가드레스
12	오산제일교회	37.136941962661815	127.06889226017749	고원식 횡단보도
13	오산시보건소	37.15818792058807	127.07683140561042	과속단속카메라
14	오산원일초등학교	37.14006663796514	127.07194036915826	보행자신호등
15	원동현대힐스테이트아파트	37.133198229405586	127.07370375386654	고원식 횡단보도
		37.13374329267566	127.07267515301915	도류화
16	원동제1어린이공원	37.131580433763745	127.07067309076014	조명시설
17	고현초등학교	37.13276520795624	127.08528797656477	과속단속카메라
		37.132639750830535	127.08533853674994	횡단보도 투광등
18	해오름어린이집	37.1325289587399	127.08468848967621	조명시설
19	오산시보건소	37.159870878696545	127.07815631773879	과속단속카메라
20	대원초등학교	37.13014723741912	127.06871348576408	주정차단속카메라

Solution 1. 모델 활용 - 예시



오산세교 복지시설의 경우, 어린이 보호구역임에도 불법 주정차가 많음
*불법 주정차는 교통사고에 영향을 미치는 주된 요인 중 하나임

따라서 이 지점에는 주정차 단속카메라 설치가 필요함

Solution 2. 선행 연구 참고 – 예시 1)



운천초등학교 정문의 경우 가드휀스가 설치되어 있지만,
비어 있는 공간에서 충분히 무단횡단이 일어날 수 있음

*무단횡단 금지시설의 설치는 사고 발생에 유의미한 감소효과가 있으므로
이 지점에는 무단횡단 금지시설의 설치가 필요함

결론 - 2 : 예시

: 안전시설물 설치 - (8)

Solution 2. 선행 연구 참고 – 예시 2)



*조명시설의 설치는 운전자의 시야 확보에 직접적인 영향을 끼침으로써
사고 발생 감소에 효과적임

몸집이 왜소한 어린이의 경우 운전자의 시야를 확보하는 것은 더욱 중요함
따라서 조명이 없는 이 지점에는 조명시설의 설치가 필요함

A large yellow diamond shape with a white border and rounded corners, positioned on the left side of the slide.

Appendix

A solid grey horizontal bar located at the bottom of the slide.

한계점

데이터

격자의 크기를 더 세분화 했다면 더 정밀한 분석이 가능했을 것

격자 단위로 분석이 진행되어 보차도 분리 여부, 도로 제한 속도 등의 변수 활용에 제한이 있었음

모델

각 변수가 사고에 어떤 영향을 미치는지에 대한 선행성을 고려한 해석이 어려움

한정된 자원(컴퓨터의 가용 소스, 분석 시간)의 문제로 더 다양한 최적화와 기법을 모델링에 적용해보지 못함

결론

선행연구를 일정부분 참고하였기 때문에, 오산시의 특징을 온전히 반영했다고 보기 어려움

추가 분석 방향

더욱 세분화되고 광범위한 변수를 수집했다면 더 높은 예측력을 가진 모델을 생성할 수 있을 것임

데이터를 시점별로 수집해 시계열 자료를 추가한다면 변수들의 선행성을 고려할 수 있을 것임

공간 분석에 최적화된 머신러닝/딥러닝 기법을 다양하게 시도하여 모델의 예측력을 높일 수 있을 것이라 기대됨

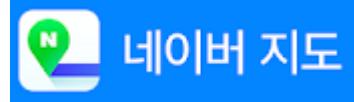
오산시의 지역적 특성을 더욱 세분화해 고려할 필요성이 있음

출처 - (1)

외부데이터

수집 데이터 셋	기준 연도	출처
36_재생산.academy_KAKAO_api.csv	2021	KAKAO DEVELOPER
18_외부.오산시_과속방지턱표준데이터.csv	2021	공공데이터포털

분석도구



kakao developers

참고논문

공간의존성에 대한 이해와 공간회귀분석의 활용, 이석환(2014)

아동 교통사고 사례분석을 통한 교통안전 문제점과 개선방안 - 아파트 단지 내 교통사고 중심으로 - , 김신혜, 임동균(2020)

교통사고 원인 분석 및 대응방안에 관한 연구, 정우진(2019)

교통사고 발생유형별 영향요인 분석을 통한 교통사고 예방대책 수립방안에 관한 연구, 배민지(2019)

도로안전시설의 사고감소효과 메타분석, 최지혜, 임준범, 이수범 (2016)

도로안전시설의 교통사고 감소효과 측정에 관한 연구, 이수범, 박규영 (2000)

교통안전시설 설치의 교통사고 발생에 대한 선행성 분석, 주일엽 (2018)

교차로에서의 도로교통안전시설물의 교통사고 감소효과도 추정, 윤여일, 이수범, 임준범, 박길수, 문정식 (2017)

Thanks for
watching

Team G.O.C