DeepLearning 모델을 이용한 대전시 교통사고 위험지역 도출 및 교통안전물의 효과 분석

어디서사고가 팀

Index

1. Intro
- 교통사고 위험 분석의 필요성
- 과제 목표
- 분석흐름도
- 제안방법의 장점 및 효과

2. Methodology
- 데이터 전처리
- 사전 연구 조사
- 변수 선택(Feature selection)
- 교통사고 수 예측모델
  - DNN 모델 소개
  - 모델 학습
  - 모델 평가
- 학습된 모델을 이용한 위험지수 정의
- 총 100개의 위험구역 산출

3. Analysis & Visualization
- 위험 지역 분석 및 시각화
  - 위험지역 클러스터링
  - 위험 지역에 영향 미치는 요인 및 영향력 분석
  - 위험 지역의 교통사고 유형/피해자 연령대/교통안전물 에 따른 사후 특징 분석
- 교통안전물 효과 분석

4. Conclusion
- 위험지역 제안
- 위험지역마다 사고 감소 효과가 큰 교통안전물 종류 제안
- 학습된 모델 활용법

- Appendix
- 추가자료목록

1. Intro

대한민국, 대전시의 교통사고 현황

- **국가 교통정책 평가지표 조사사업** (한국교통연구원, 2020):
  - 2018년 한 해 동안 GDP의 2.2%인 41조원의 사회적 비용이 도로교통사고로 인해 발생
  - 사망자수 및 중상자수가 감소했음에도 불구하고 경상자수와 부상신고자수의 증가로 인해 전년도에 비해 도로교통사고비용이 약 4% 증가
  - 1,228,129건의 교통사고 발생,3,657명 사망, 1,935,008 명 부상
  - 사상자의 물리적 손실비용(생산손실, 의료비, 물적피해 등)은 약 22조, 정신적 고통비용은 약 19조원으로 추정
  - 다른나라의 GDP 대비 도로교통사고비용 비교 시 미국 1.85%(2010년), 일본 1.35% (2009년), 영국 1.81%(2017년)로 우리나라의 2.20%(2018년)는 주요국 대비 여전히 높은 수준

- 대전시는 2018년 대비 2019년에 (도로교통공단,2020)
  - 사망자 수는 17.2% 감소하였으나, 전체 사고건수는 10.4% 증가
  - 자전거 가해자사고건수는 68.8% 증가, 사망자 수는 2명에서 4명으로 100% 증가.
  - 비슷한 인구수를 가지는 광주시에 비해 약 5% 적은 사고 건수.

사상자구분(명)	2017년	2018년	사상자수 변화	
사망자	4,185	3,781	-404	(-9.7%)
중상자	96,810	91,985	-4,825	(-5.0%)
경상자	581,589	639,999	+58,410	(+10.0%)
부상신고자	1,124,926	1,203,024	+78,098	(+6.9%)

표1. 2017년, 2018년 도로교통사고 사사장수 비교(한국교통연구원)

〈표〉 시도별 도로교통사고비용('18년)

(단위: 천 원, 건, 인, 천 원/인)

시도	사고비용	사고건수	인구	인구 1인당 비용
서울	4,997,423,712	217,598	9,673,936	517
부산	1,796,422,587	71,028	3,395,278	529
대구	1,653,448,493	66,549	2,444,412	676
인천	1,510,209,908	66,614	2,936,117	514
광주	976,019,346	42,417	1,490,092	655
대전	1,000,781,008	40,698	1,511,214	662
울산	689,216,533	26,919	1,150,116	599
세종	126,013,611	4,626	312,374	403
경기	7,718,607,944	307,909	13,103,188	589
강원	1,154,137,065	36,330	1,520,391	759
충북	1,257,968,688	35,081	1,620,935	776
충남	1,732,666,341	51,784	2,181,416	794
전북	1,291,759,804	43,304	1,818,157	710
전남	1,355,354,739	38,075	1,790,352	757
경북	2,065,103,459	61,017	2,672,902	773
경남	2,192,949,755	67,995	3,350,350	655
제주	517,366,570	16,563	658,282	786
계	32,035,449,563	1,194,507	51,629,512	620

자료: TAAS(교통사고분석시스템), 통합DB 기준

교통사고 위험 분석의 필요성 및 목표

- 안전체계(Safe System)개념 (1990s)
  - '사람은 누구나 실수할 수 있고, 교통사고가 나더라도 사람이 죽거나 다치게 해서는 안된다'
  - 사람을 둘러싼 도로환경의 개선을 중요시
  - 교통사교는 안전체계를 통해 사람과 시스템이 위험을 부담할 때, 효과적으로 예방 가능

- **기존 교통사고 위험 연구에 따르면 지역마다 교통사고 위험 분석이 필요하다. 그 이유는 다음과 같다.**
  - 교통사고의 위험은 지역에 따라 다름.[1]
  - 교통사고 위험은 운전자, 차량, 도로, 환경 등의 요인으로 설명 가능한데[2], 지역별로 이 요인들이 다르므로 지역별로 교통사고의 위험도가 다름.

목표

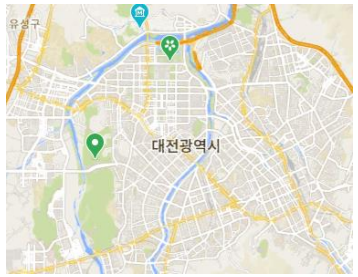
- 주어진 3년간의(2017-2019) 대전시 데이터(교통사고 데이터, 교통안전시설물, 거주인구, 교통량, 건물연면적, 도로 연장 데이터 등)를 활용해
  - 연령대별 교통사고 유형, 교통안전물과 사고유형 사이의 관계, 사고 유형별 분포, 도로혼잡빈도와 사고수의 관계 등의 데이터 분석
  - DNN을 이용한 구역별 교통사고 발생횟수 예측 모델 학습
  - 교통사고 예측치가 실제보다 높은 지역을 위험지역으로 제안.
  - 군집화 및 XAI 모델(LIME)을 이용하여, 각 위험지역별 요인 분석
  - 위험구역별 사고감소효과가 가장 좋은 교통안전물의 종류 및 그 효과 분석

- 이 과제의 목적은 각종 데이터를 분석하고, 이를 이용한 Deep learning model을 학습하여, 도로교통안전 분야의 의사결정에 지표 또는 자료로 활용될 수 있도록 교통사고 위험지역을 예측 탐지하고, 교통안전물의 추가 설치 효과가 높은 구역을 제안하는 데에 있다.

[1] G. C. Taylor, Use of spline functions for premium rating by geographic area, ASTIN Bulletin: The Journal of the IAA, 19(1) (1989), pp. 91-122.
[2] G. Zhang, K. K.W. Yau, AND G. Chen, Risk factors associated with traffic violations and accident severity in China, Accident Analysis and Prevention, 59 35 (2013), pp. 18-25.

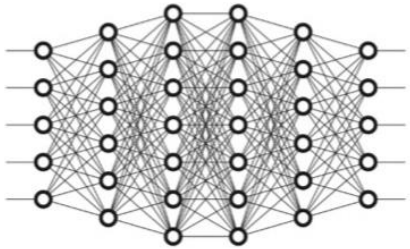
분석 흐름도

1. 데이터 수집 및 전처리



- 대전시 데이터셋 21개 수집
- 데이터 분석
- 데이터 전처리
- 변수 선정
- 최종 18개 변수 이용

2. 모델 학습

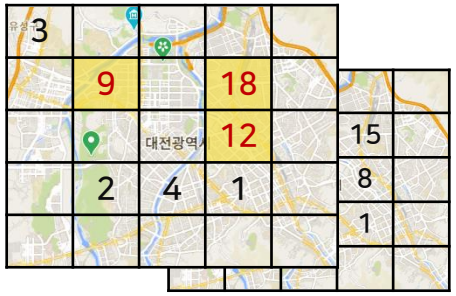


1. 구역당 교통사고횟수를 예측하는 Deep Neural Network(DNN) 모델 학습

2. 모델 평가

- Linear, Ridge, Lasso regression, Decision Tree, Random Forest 모델과 비교

3.1.사고위험지역 추출



1. **DNN 사고위험지수** 정의 및 제안

(DNN사고위험지수)

=(예측 사고횟수) - (실제사고횟수)

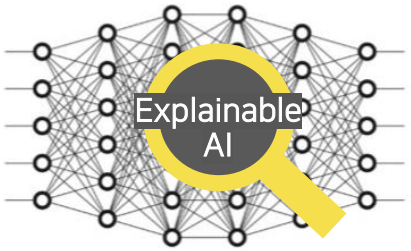
2. 정확도를 높이기 위한 사후처리 진행

3. 사고위험지수가 높은 사고 위험 지역 제안

4.2. 교통안전물 추가 설치 효과 분석

학습된 DNN 모델을 활용해 교통안전물의 추가 설치 효과가 큰 구역과 추가 설치할 교통안전물 종류 제안

4.1. 사후 분석 및 시각화



1. 위험지역의 특성을 파악하기 위해, 위험지역의 군집화(K-means clustering)를 총 3가지 유형으로 진행

2. XAI(Explainable AI, 설명가능한 AI)을 통해 **사고위험지역의 원인 분석**

제안 방법의 장점 및 효과

1. 최신 기술인 딥러닝(Deep learning) 방법 중 하나인 DNN(Deep Neural Network) 모델을 사용하여 학습하였기 때문에, 기존 모델에 비해 **예측 정확도 향상**됨.

2. 이러한 딥러닝 모델은 추론 과정을 쉽게 파악하기 힘든 Black box 모델이라는 단점이 있으나, 최근 활발히 연구 중인 **설명 가능한 AI(Explainable AI) 기술을 접목**하여 이 문제점을 해결하고, **사고 위험지역으로 예측한 원인을 파악**할 수 있음.

3. 딥러닝 모델은 사람이 학습하는 것보다 더 많은 지식을 빠르게 학습하기 때문에, 간과하던 변수들의 중요성을 인지하고 **각 변수들의 중요성을 수치화 할 수 있음**.

4. 데이터를 통해 학습된 특정 구역의 **교통안전물 개수 증감에 따른 발생사고 증감을 수치화**할 수 있음.

5. 시간이 지남에 따라 **추가되는 데이터를 넣어 쉽게 학습**할 수 있음.

6. 추후 데이터 양이 많아질수록 위험지역 예측 정확도가 증가함.

2. Methodology

기존 교통사고 위험 연구 조사

* 과제의 목적이 교통사고 위험 지역을 제안하는 것이기 때문에, 데이터를 통해 교통사고 위험지역을 예측 후 제안하고자 하였고, 이에 따라 관련 연구를 조사하였다.

- 교통사고 위험 연구는 크게 다음과 같은 세가지로 분류 가능
  - 1. 교통사고 발생 특성에 관한 연구: 교통사고 발생을 설명할 수 있는 요인 찾고, 교통사고의 빈도 또는 심도 위험을 설명
  - 2. 교통사고 발생 빈도 예측 연구: 교통사고의 발생 여부 및 발생 횟수 예측
  - 3. 교통사고의 심도 예측 연구: 교통사고가 일어날 경우에 그 사고의 규모 또는 피해액 예측

  - 밑에 슬라이드 노트 내용 보고 내용 추가!!

* 참고 문헌 목록은 슬라이드 맨 뒤 APPENDIX에 첨부

해외 교통사고 위험 연구 조사

- 해외의 경우, 과거에는 2000건 미만의 데이터를 이용해 단순 회귀 모델, 분류 모델등의 기계학습 모델을 이용하여 교통사고 위험도를 예측
- 최근에는 기존 모델에 비해 정확도가 높은 최신 AI 기술인 CNN, LSTM 등의 여러가지 Deep Learning 모델을 이용하여 위험도를 예측.

(과거) 단순 기계학습 모델 사용

- 스페인 그라나다 지방 도로 교통사고 데이터(2003-2009) 1801건 을 이용한 다수의 의사결정 나무 이용하여 규칙을 생성하는 모형 제안 [3]
- 군집 찾기 알고리즘으로 미국 I-190 도로 교통사고 데이터(2008-2012) 999건을 군집화하여 데이터의 내재적 이질성을 줄인 후 연관규칙 학습 알고리즘을 각 군집에 적용하여 교통사고 빈도를 설명하는 요인 탐색[4]
- Classification and Regression Tree(CART) 모델과 대만 1번 고속도로 교통사고 데이터 (2001-2002) 1075건 이용하여 교통사고 빈도 위험 예측[5]
- 푸아송, Negative Binomial, Negative Multinomial 회귀 모델을 이용하여 이탈리아 교통사고 데이터(1999-2003) 1916건을 바탕으로 다차선 도로상에서의 교통사고 발생 빈도 예측[6]
- Freequent Pattern Tree 라는 변수 선택 방법론을 제안하고 미국 버지니아 I-64 고속도로 교통사고 데이터(2005) 344건을 이용해 k-Nearest Neighbor(KNN), 베이지안 네트워크 모델을 구축하여 실시간 교통사고 발생 위험 예측 [7]
- Convolutional Neural Network(CNN)을 기반으로 한 새 모델을 제안하고 미국 I-15고속도로 교통사고 데이터 250건데 적용하여 사고 발생 예측[8]
- 베이지안 네트워크 모델을 이용해 스페인 그라나다 지방 고속도로 교통사고 데이터(2003-2005) 1536건에 적용하여 교통사고 심도 예측[9]
- Multinomial Logit, Nearest Neighbor Classification, Support Vector Machine, Random Forest 의 네가지 모델을 이용하여 미국 네브레스카 교통사고 데이터 (2012-2015) 68448 건을 바탕으로 교통사고 심도 예측[10]

(최근 연구) 딥러닝 모델 사용

- Long Short-Term Memory(LSTM) 모형을 기반으로 시공간적 상관관계를 고려하는 모형인 Traffic Accident Risk Prediction Method based on LSTM(TARPML)을 그림 2.1과 같이 제안하고 중국 베 이징 교통사고 데이터(2016-2017) 2,222,548건에 적용하여 교통사고 빈도 위험을 예측[11]
- 기상 정보, 도로 정보, 인공위성 사진 등의 이질성 데이터를 고려하도록 Convolutional Long Short-Term Memory(ConvLSTM) 기반으로 Hetero-ConvLSTM 모형을 제안하고 미국 아이오와 교통사고 데이터(2006-2013)에 적용해 교통사고 발생 을 예측 [12]
- Attention 기반 Residual Network(ResNet) 모형으로 미국 뉴 욕 지역에 대해 교통사고 데이터(2017)와 교통 관련 다른 분야의 데이터를 함께 학습 하여 교통사고 발생을 예측[13]
- Convolutional Neural Network(CNN), Long Short-Term Memory(LSTM), Convolutional Long Short-Term Memory(ConvLSTM)를 기반으로 Spatiotemporal Convolutional Long Short-Term Memory Network(STCL-Net) 모형을 제안하고 도시 내에서 단기간의 교통사고 심도 10 위험을 예측. 미국 맨해튼 지역 교통사고 데이터(2015) 53,354건에 적용[14]

* 참고 문헌 목록은 슬라이드 맨 뒤 APPENDIX에 첨부

국내 교통사고 위험 연구 조사

- 국내의 경우, 해외에 비해 교통사고 위험 연구 조사의 수가 매우 적음.
- 과거 국내 연구의 경우 해외와 마찬가지로 쉽게 학습이 가능하지만 성능이 낮은 기계학습 모델을 통해 교통사고 위험 예측.
- 최근 국내 연구의 경우 딥러닝 모델을 한정적인 지역에 사용하거나, 해외 데이터에 적용.

(과거) 단순 기계학습 모델 사용

- LightGBM[15], 다중선형회귀분석[18] 등이 쉽게 학습이 가능하지만 성능이 낮은 기계학습 모델을 이용
- 다중 결합 예측 알고리즘(MAPA)의 시계열 분석 방법을 이용[17]
- 단순히 사고가 일어난 지역인지 아닌지를 판별하는 모델[16]을 이용

(최근) 한정된 딥러닝 모델 사용

- 딥러닝 모델을 이용한 전체지역이 아닌 고속도로 위의 사고 예측[20]의 특정 구역 예측에서의 연구.
- 넓은 범위에 적용된 경우 국내 데이터가 아닌 해외 데이터를 이용해 연구가 진행[21,22]되어, 국내 데이터에 적용 여부를 확인하기 어려움.

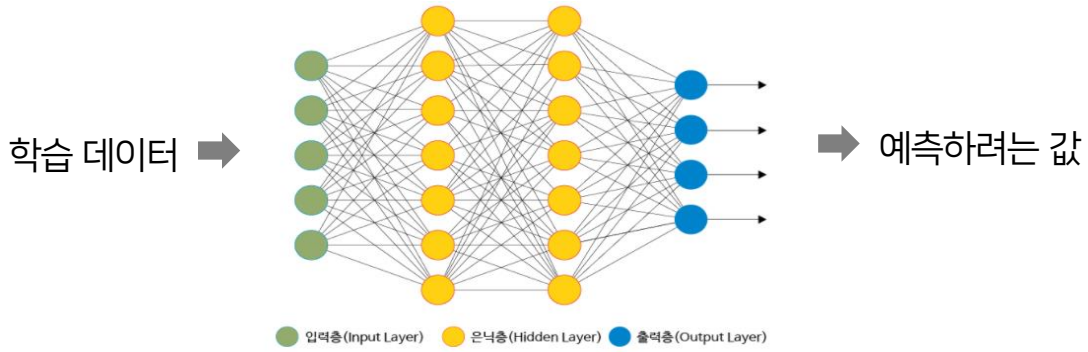
따라서, 기존 교통사고 위험 예측 연구에서 성능이 가장 좋았던 **딥러닝 모델 채택**

의의

- (조사한바에 따르면) 딥러닝 모델을 이용해 국내 시 범위의 교통사고 횟수를 예측한 최초의 연구
- 대전시 데이터로 학습되어, 대전시 교통사고 횟수 예측에 적합한 모델 생성 가능

* 참고 문헌 목록은 슬라이드 맨 뒤 APPENDIX에 첨부

딥러닝이란 무엇인가

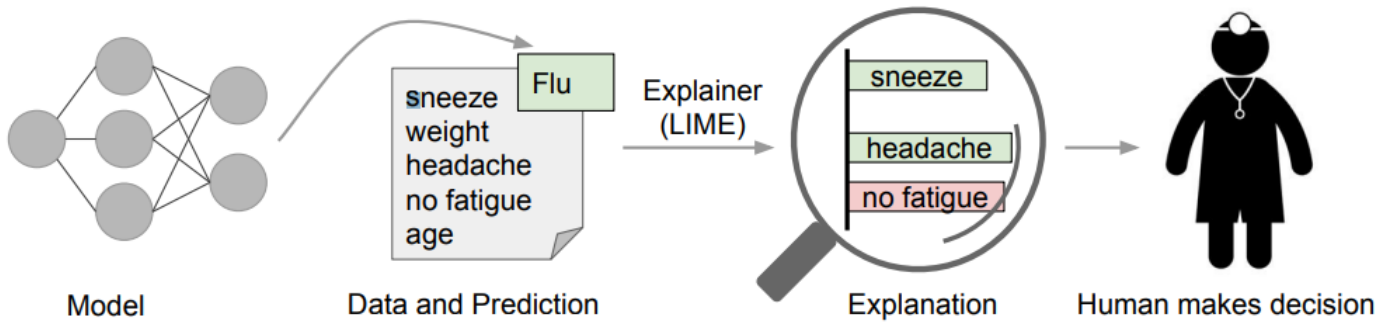


- 컴퓨터가 자동으로 대규모 데이터에서의 중요한 패턴 및 규칙을 학습하고, 이를 토대로 의사결정이나 예측을 수행하는 기술.
- 여러 층을 가진 인공신경망(Artificial Neural Network)을 사용하여 학습을 수행
- 3개 이상의 레이어를 가지는 경우 깊은신경망(Deep Neural Network, DNN) 모델이라고 정의.
- 어떤 변수를 추출할 지 선택하지 않아도, 컴퓨터가 중요한 변수를 스스로 선택하여 학습
  - 변수 선택(Feature selection)의 중요성 감소- 사람의 bias(선입견)의 영향력이 적음
  - 데이터 하나마다 각 변수의 영향력을 다르게 학습,
    - 예) 모든 구역에서 교통량이 많다고 사고가 많이 일어나는 것이 아니라 구역의 특징에 따라 교통량이 사고횟수에 얼만큼 어떠한 영향력을 지니는지를 다르게 학습.

머신러닝과의 차이점

- 머신러닝(기계학습): 여러 레이어를 사용하지 않는 선형 회귀분석, 로지스틱 회귀분석, k-최근접이웃알고리즘 등의 회귀분석, 클러스터링 방법.
- 딥러닝은 머신러닝의 한 종류이며, 머신러닝은 주로 한두단계의 레이어를 사용한다면, 딥러닝은 여러 개의 레이어를 사용하여 학습한다.
- 머신러닝에서는 학습하려는 데이터의 여러 특징 중에서 어떤 특징을 추출할지를 사람이 직접 분석하고 판단해야 했다면, 딥러닝에서는 컴퓨터가 자동으로 추출하여 학습.

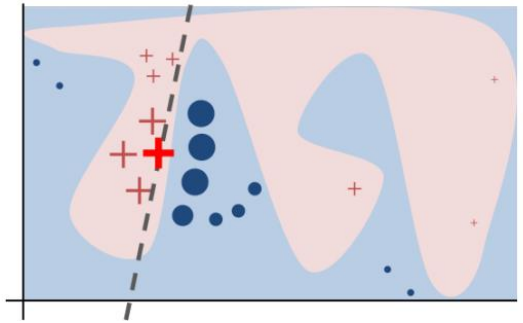
Explainable AI(XAI, 설명가능한 AI)



[23]

- **블랙박스(black box)인 딥러닝 모델**
  - 상호 연결된 노드의 아주 복잡한 구조로 되어있으므로, 특정 결론에 이르는 과정이 투명하지 않고 모호함.
  - 모델이 예측한 결과물의 근거를 파악하기 어려움.
- 어떻게 예측하였는지 해석이 불가능한 딥러닝 모델에서 예측한 사항의 근거를 사람이 이해할 수 있도록 설명해주는 기술.

LIME(Local Interpretable model-agnostic explanations)



- 임의의 블랙박스 모델을 이미 설명이 가능한 데이터 주변에서 희소 선형 결합을 통해 국부적으로 설명 가능하도록 개발된 방법[23]

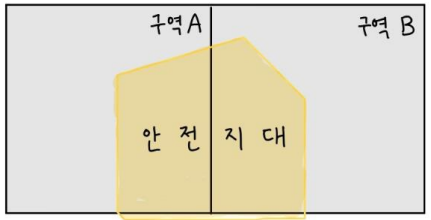
예) 왼쪽 그림에서 데이터를 색으로 분류를 한다고 할 때, 모든 데이터를 분류하는 모델을 한번에 설명하기는 어렵다. 하지만 빨간 십자가 데이터를 해석한다고 할 때, LIME은 그 근처 데이터를 이용해 학습하고 그 결과 점선으로 분류 가능하며 쉽고 비교적 정확하게 해석이 가능하다.

Python LIME library를 통해 구현가능.

데이터 전처리

- 위치
  - 데이터에 따라 위치가 WGS84좌표계로 표현된 경우와 gid(격자고유 ID)로 표현된 두가지 경우가 있으므로, 좀 더 큰 범위인(100m*100m) gid로 통일
- 교통 안전물
  - 점(point)이 아닌 면(polygon) 형태로 구성된 경우가 있고, 해당 경우에는 여러 구역(gid)에 겹쳐 있을 수 있으므로 겹친 모든 gid 정보가 필요
  - [문제점] 모든 gid를 탐색하려면 O(n^2) 시간이 걸림- 주어진 자원으로 수십시간 이상 걸림.
  - [해결] STRtree 데이터 구조를 이용하여 O(n) 시간에 탐색할 수 있었음.

그림1. 교통안전물 전처리



- 구역당 도로길이
  - 구역당 교통량을 계산하기 위해 해당 구역을 교차하는 모든 도로의 길이의 합을 계산
    - 그림과 같이 교차하는 도로의 전체 길이가 아닌 정확히 모든 겹치는 구역의 도로 길이를 계산하기 위해 python shapely library 이용
    - 효율적인 계산을 위해 STRtree 데이터구조 사용
- 사고 유형
  - ' 차량단독' 사고 유형의 경우, 전체 차량단독 사고가 3%로 매우 적으므로, 세부 유형으로 나누지 않음
- 피해자 연령대
  - 피해자별 사고 유형에 따라 (10대미만, 10대), (20대), (30~50대), 60대, (70~90대) 로 그룹
- 교통량
  - 교통량을 총 4가지로 정의.(자세한 정의와 처리방법은 appendix에 첨부)

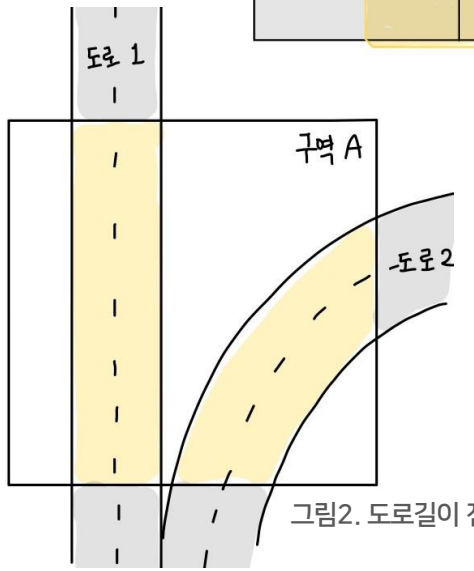


그림2. 도로길이 전처리

* 1.ipynb 에서 진행

1. 연령대별 사고 유형 분석 및 그룹화

피해운전자 연령대	10대	10대미만	20대	30대	40대	50대	60대	70대	80대	90대
피해운전자 연령대										
10대	1.000000	0.798919	0.177053	-0.828747	-0.925136	-0.650123	0.109361	0.681118	0.790315	0.280518
10대미만	0.798919	1.000000	-0.174337	-0.851277	-0.838618	-0.471215	0.301344	0.825228	0.927328	0.545919
20대	0.177053	-0.174337	1.000000	0.001747	-0.317013	-0.714806	-0.842762	0.054225	0.043637	-0.103168
30대	-0.828747	-0.851277	0.001747	1.000000	0.869212	0.517054	-0.365112	-0.906937	-0.851826	-0.588795
40대	-0.925136	-0.838618	-0.317013	0.869212	1.000000	0.763502	0.036974	-0.809197	-0.907671	-0.478632
50대	-0.650123	-0.471215	-0.714806	0.517054	0.763502	1.000000	0.470348	-0.626311	-0.525054	-0.347379
60대	0.109361	0.301344	-0.842762	-0.365112	0.036974	0.470348	1.000000	0.221303	0.127468	0.302973
70대	0.681118	0.825228	0.054225	-0.906937	-0.809197	-0.626311	0.221303	1.000000	0.780014	0.719981
80대	0.790315	0.927328	0.043637	-0.851826	-0.907671	-0.525054	0.127468	0.780014	1.000000	0.579450
90대	0.280518	0.545919	-0.103168	-0.588795	-0.478632	-0.347379	0.302973	0.719981	0.579450	1.000000

표1. 사고유형 비율을 이용한 피해 운전자별 연령대 상관 분석

가해운전자 연령대	10대	10대미만	20대	30대	40대	50대	60대	70대	80대	90대
가해운전자 연령대										
10대	1.000000	-0.033743	0.082103	-0.200616	0.138091	-0.643480	-0.543967	0.378447	0.056921	-0.167808
10대미만	-0.033743	1.000000	0.041179	-0.090012	-0.079852	0.090991	0.168835	-0.133578	-0.014147	-0.076603
20대	0.082103	0.041179	1.000000	0.530771	-0.577305	-0.529149	-0.528866	-0.016388	0.018892	-0.041048
30대	-0.200616	-0.090012	0.530771	1.000000	0.100583	-0.467736	-0.499176	-0.600386	-0.029094	0.045994
40대	0.138091	-0.079852	-0.577305	0.100583	1.000000	0.068560	-0.242939	-0.566128	0.013925	0.087507
50대	-0.643480	0.090991	-0.529149	-0.467736	0.068560	1.000000	0.575889	-0.170967	0.081028	0.055344
60대	-0.543967	0.168835	-0.528866	-0.499176	-0.242939	0.575889	1.000000	0.271791	0.029151	0.072993
70대	0.378447	-0.133578	-0.016388	-0.600386	-0.566128	-0.170967	0.271791	1.000000	-0.310080	-0.079754
80대	0.056921	-0.014147	0.018892	-0.029094	0.013925	0.081028	0.029151	-0.310080	1.000000	-0.071526
90대	-0.167808	-0.076603	-0.041048	0.045994	0.087507	0.055344	0.072993	-0.079754	-0.071526	1.000000

표2. 사고유형 비율을 이용한 가해 운전자별 연령대 상관 분석

피해자 연령대별 사고유형 분석

- 상관계수 분석을 통해 사고 유형이 비슷한 피해자 나이대 그룹화
- (10대미만, 10대), (20대), (30~50대), (60대), (70~90대)의 5개의 연령대로 그룹화
- 90대는 70대,80대와 상관계수가 0.71, 0.57 로 약한 상관관계를 가지지만 그 수가 매우 적어 같은 집단으로 처리.

가해자 연령대별 사고유형 분석

- 사고 유형이 비슷한 가해자 연령대는 없음 (상관계수 0.7 미만)
- 연령대별 가해자의 사고유형은 전부 다름.
- 가해자의 연령대별 그룹화 진행 안함.

Feature Selection (변수 선택)

* 모든 변수는 구역 단위(100m*100m)로 계산

변수 추가

기존 연구 결과에 따라 교통사고발생에 영향을 주는 변수 추가

- 인구
  - 전체 인구수
  - 유소년 인구수
  - 생산가능 인구수
  - 고령 인구수
- 도로
  - 교차 도로 개수
  - 교차 도로 길이
  - 교차로 개수
  - 속성 변화점 개수
  - 도로시설물 개수
  - ic 및 jc 개수
- 건물: 건물 연면적
- 차량: 자동차 등록 대수
- 교통안전물:
  - 신호등 개수
  - 안전지대 개수
  - 횡단보도 개수
  - 도로속도표시 개수,
  - 정차금지지대 개수
  - 교통cctv 개수
  - 교통안전표시 개수
  - 중앙분리대 개수
- 교통량: 교통량1, 교통량2, 교통량3, 교통량4 를 정의

변수 제거

상관관계가 높은 변수 제거 혹은 변수 혼합을 통한 새로운 변수 추가

- 총인구수와 유소년 인구수, 생산가능 인구수, 고령 인구수 사이의 상관계수가 각각 0.90, 0.99,0.86 로 서로 높은 상관관계

-> '총인구수'를 제외한 모든 연령별 인구수 변수 제거

- 도로개수와 교차도로길이가 0.81의 높은 상관계수 가지므로 서로 높은 상관관계에 있음.

-> '도로개수' 변수 제거

- 교통량2, 교통량3, 교통량4 가 서로 0.75 이상의 높은 상관계수를 가짐

-> 교통량2, 교통량3 제거

- 보행등, 차량등 개수가 0.85의 높은 상관계수 가짐

-> 두 변수의 평균값을 가지는 '신호등_개수' 변수 추가 후 두 변수 제거.

최종 독립 변수 목록

변수 선택 후, 모델 학습에 사용된 최종 독립 변수는 다음과 같다.

(단위: 각 gid 당 명/개수)

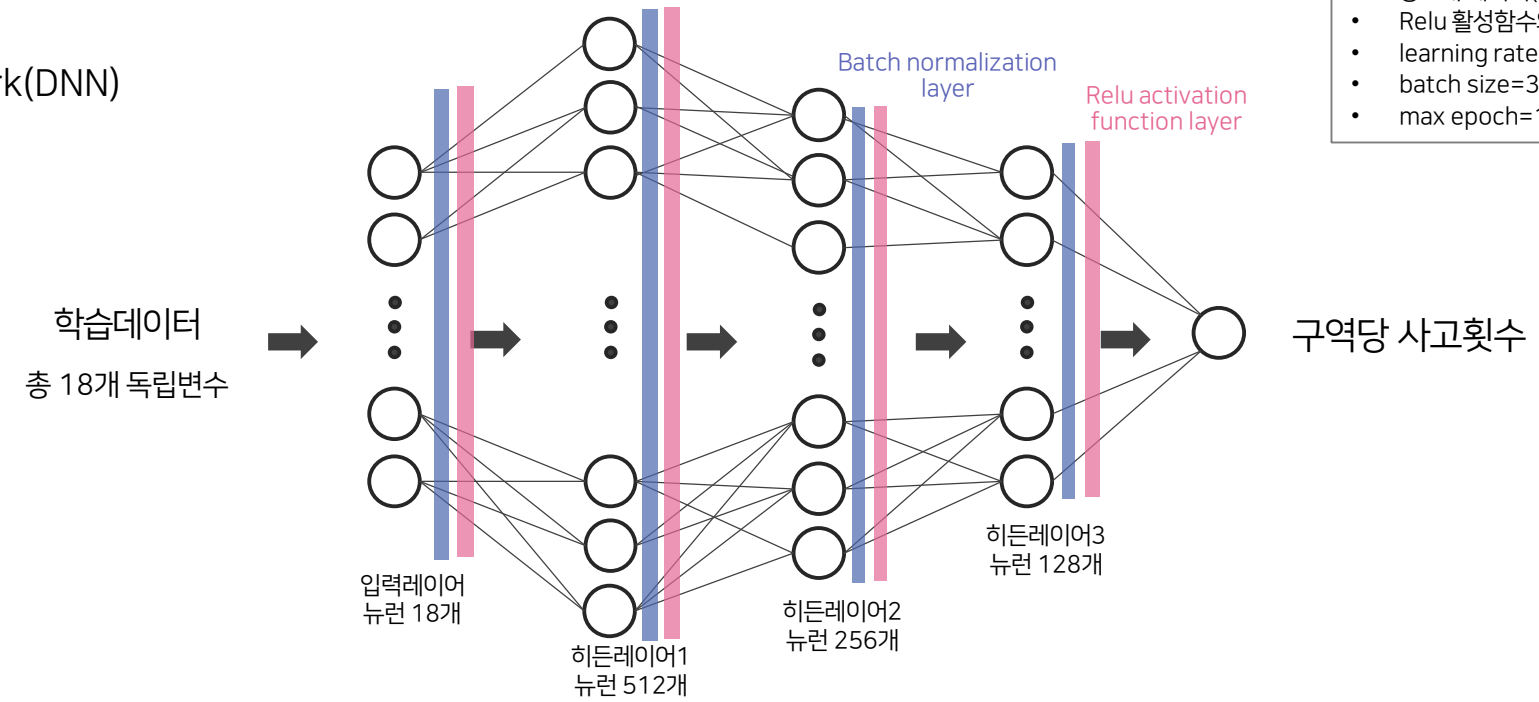
분류	상세 변수명	정의	추출한 데이터 파일
인구	전체인구수	해당 구역에 살고 있는 전체 인구 수	12.대전광역시_인구정보(총인구).geojson
도로	도로연장(교차도로길이)	해당 구역에 포함된 모든 도로의 길이	19.대전광역시_상세도로망(2018).geojson
	교차로	해당 구역에 포함된 모든 교차로 개수	18.대전광역시_교통노드(2018).geojson
	속성변화점	해당 구역에 포함된 모든 속성변화점 개수	18.대전광역시_교통노드(2018).geojson
	도로시설물	해당 구역에 포함된 모든 도로시설물 개수	18.대전광역시_교통노드(2018).geojson
	Ic및jc	해당 구역에 포함된 모든 ic 및 jc 개수	18.대전광역시_교통노드(2018).geojson
건물	건물연면적	해당 구역의 건물 총 연면적(값이 없을 경우 0 처리)	24.대전광역시_건물연면적_격자.geojson
차량	자동차등록대수 (cars_cnt)	해당 구역의 차량등록현황	30.대전광역시_차량등록현황_격자.geojson
교통량	교통량1	(해당 구역과 교차하는 도로 길이) * (각 도로의 혼잡빈도강도) 의 총합	21.대전광역시_평일_일별_혼잡빈도강도(2018).csv
	교통량4	해당 구역을 교차하는 모든 도로의 혼잡빈도강도 평균값	21.대전광역시_평일_일별_혼잡빈도강도(2018).csv
교통안전물	신호등 개수	해당 구역의 신호등(보행등 및 차량등) 개수	3.대전광역시_신호등(보행등).geojson, 4.대전광역시_신호등(차량등).geojson
	안전지대 개수	해당 구역의 안전지대 개수	5.대전광역시_안전지대.geojson
	횡단보도 개수	해당 구역의 횡단보도 개수	6.대전광역시_횡단보도.geojson
	도로속도표시 개수	해당 구역의 도로속도표시 개수	7.대전광역시_도로속도표시.geojson
	정차금지지대 개수	해당 구역의 정차금지지대 개수	8.대전광역시_정차금지지대.geojson
	교통안전표지 개수	해당 구역의 교통안전표지 개수	9.대전광역시_교통안전표지.geojson
	교통CCTV 개수	해당 구역의 교통CCTV 개수	10.대전광역시_교통CCTV.geojson
	중앙분리대 개수	해당 구역의 교통중앙분리대 개수	31.대전시_중앙분리대.geojson

* 2_2.ipynb 에서 진행

DNN 교통사고횟수 예측모델 학습

사용한 Model Parameter 정보
- 총 3개 레이어(각각 뉴런 512,256,128개로 구성)
- Relu 활성함수와 각 레이어 사이 배치 정규화 진행
- learning rate=$8 * 10^{-6}$, Adam optimizer
- batch size=32
- max epoch=150, early stopping (patience=8)

Deep Neural Network(DNN)
모델구조



1. train data(90%), test data(10%) 로 분리 후, train data만으로 학습시키고 test data로 모델 성능 평가
2. 데이터 스케일링을 위해 각 feature마다 평균이 0이고 분산이 1인 표준정규분포로 변환하는 **표준화**(standardization) 진행.
3. 레이어 개수, 뉴런 개수, 활성함수, 배치정규화 진행유무, dropout rate 등의 여러 **매개변수를 변경하며 최적 성능 모델을 탐색**.

- 최적 성능 모델은' 512_256_128_9e_6_d09_batch32_3273' 로 저장하여 로드가능.
- 추가로 총 교통사고 발생횟수가 아닌 피해자 연령대별 사고횟수에 따라서도 모델 학습을 진행하고자 하였으나, 데이터의 빈도수가 너무 적어져 오차가 너무 커져 학습이 불가능했다.
- Python keras library 이용

DNN 교통사고횟수 예측모델 성능 평가

* 2_2.ipynb 에서 진행

* Test data에 대한 평가 결과
* 비교 모델들과 DNN 모델의 상세 평가 수치 및 train data에 대한 수치는 appendix에 첨부.

	DNN	Linear 회귀	Ridge 회귀	Lasso 회귀	Decision Tree	Random Forest
실제 사고 지역에 대해 무사고로 예측할 확률	**11.9%**	36.2%	36.2%	34.5%	40.7%	31.3%
실제 무사고 지역에 대해 무사고로 예측할 확률(specificity)	91.3%	96.2%	96.2%	96.1%	95.5%	96.3%
실제 사고 지역에 대해 사고 지역이라고 예측할 확률(recall)	**88.0%**	63.8%	63.8%	65.4%	59.2%	68.6%
사고 예측 지역이 실제 사고 지역일 확률(precision)	**76.6%**	67.6%	67.6%	67.5%	61.9%	69.5%
실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률	95.5%	97.7%	97.7%	97.2%	69.4%	100%
실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률	96.2%	99.1%	99.1%	98.7%	73.5%	100%
실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률	**100.0%**	100%	100%	100%	89.1%	100%
사고가 2회 이상 발생할 것이라고 예측한 지역에 대해 실제로 그럴 확률	**68.4%**	55.2%	55.2%	54.9%	61.9%	61.3%
사고가 10회 이상 발생할 것이라고 예측한 지역에 대해 실제로 그럴 확률	**100.0%**	47.4%	47.4%	44.4%	24.4%	38.4%

* 비교 모델들은 Cross-validation(n=10)을 통해 최적의 parameter값을 가지도록 설정.

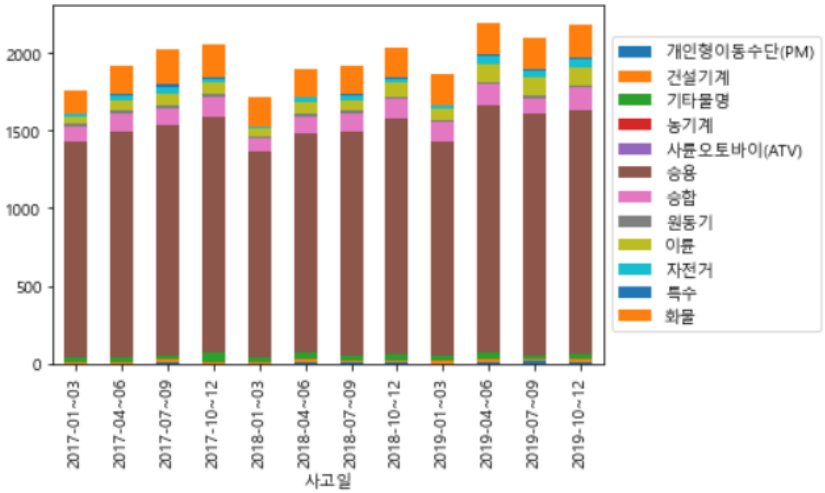
* 사고지역을 양성, 무사고지역을 음성

- 제안하는 DNN 모델의 성능을 평가하기 위해, 자주 사용되는 회귀 모델들(Linear, Ridge, Lasso regression, Decision Tree, Random Forest)과 비교.
- 데이터 상 무사고 구역이 많기(89%) 때문에, 단순히 정확도로 모델의 성능을 평가할 수 없음 (전부 무사고라고 해도 89%의 정확도를 보일 수 있기 때문)
- 따라서, 예측 사고 구역이 실제 사고 구역인지 판단하는 정밀도(Precision)와 실제 사고 구역을 사고 구역이라고 판단하는 재현율(Recall)이 중요.
- DNN 모델의 경우 비교 모델들과 비교했을 때, 정밀도와 재현율에서 가장 좋은 성능을 보임.
- 또한 실제 사고 지역을 무사고로 예측할 확률이 가장 적음.
- 예측하려는 사고 횟수의 밀도가 낮기 때문에, 사고횟수를 정확히 맞추는 것보다 사고가 많이 발생한 지역을 많이 발생했다고 예측하는 것이 중요.
    - DNN 모델의 경우, 사고가 10회 이상 발생할 것이라고 예측한 구역에 대해 실제로 그럴 확률이 100%로 다사고 지역에 대해 높은 예측 확률을 보임.

3. Analysis & Visualization

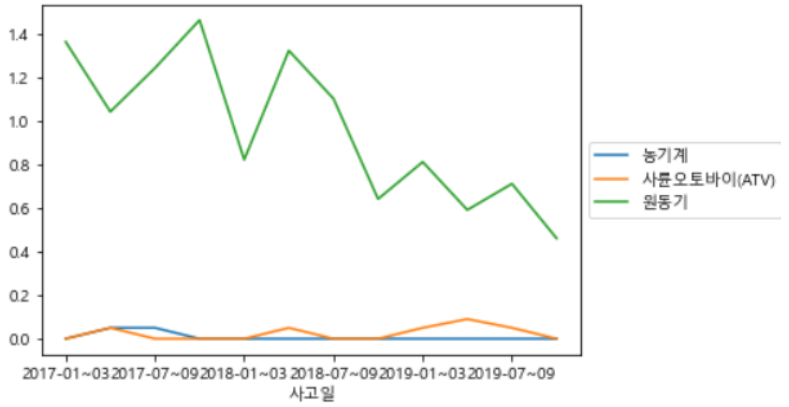
1. 교통사고 추이 분석

그래프1. 기간별 사고수 및 가해운전자 차종에 따른 사고수(3개월 단위)

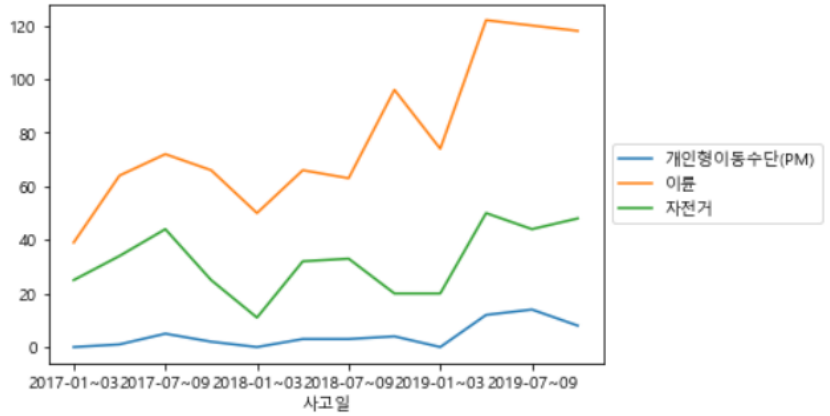


1. 대전시 전체 사고수는 지난 3년간 (2017~2019) 큰 증감이 없음.

그래프2. 기간별 가해운전자 차종 별 사고수 추이 그래프 (3개월 단위)



2. 농기계, 사륜오토바이, 원동기가 발생시킨 사고 횟수는 지난 3년간 감소하는 경향이 있음.



3. 개인형이동수단, 이륜, 자전거가 발생시킨 사고 횟수는 지난 3년간 증가.

2017년부터 2019년까지 대전시 전체 교통사고 수는 큰 증감이 없으나(+7%), 전동킥보드 및 개인용 자전거를 포함하는 이륜차, 자전거, 개인형 운송수단의 2019년 교통사고 수가 2017년에 비해 약 67% 증가

제목	모든 차종	이륜차	자전거	개인형운송수단(PM)
2017년	7759	241	128	8
2018년	7552 (-3%)	275 (14%)	96 (-15%)	10 (25%)
2019년	8341 (7%)	434 (80%)	162 (26%)	34 (325%)

표1. 해당차종이 가해한 교통사고 발생 횟수(2017년 대비 증가율)

* 1.ipynb 에서 진행

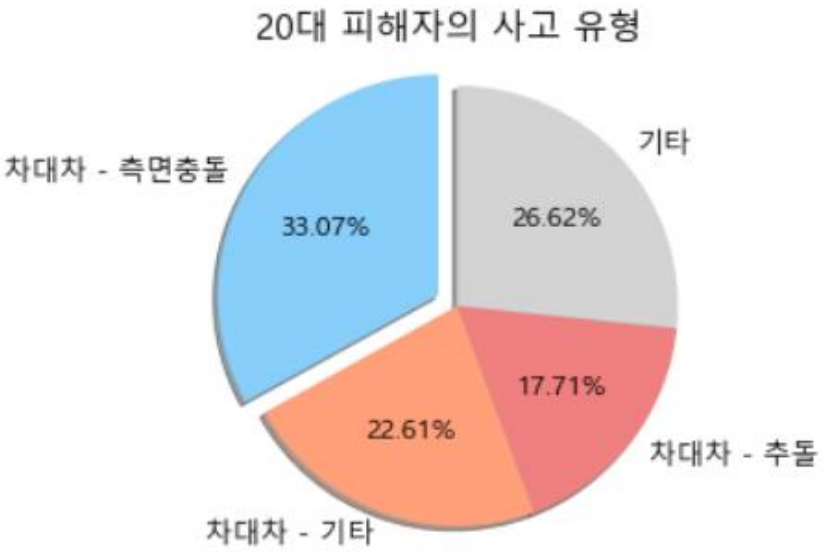
2. 피해자 연령대별 사고 유형 시각화

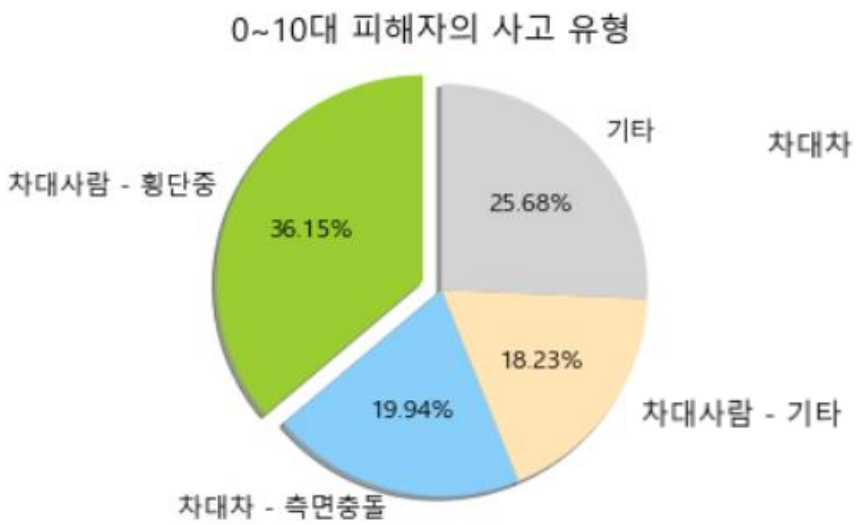
* 피해자 연령대별 사고유형 상세 비율과 추가로 진행한 사고유형별 피해자 연령대 사이의 상관관계 분석은 APPENDIX 2 에 첨부.

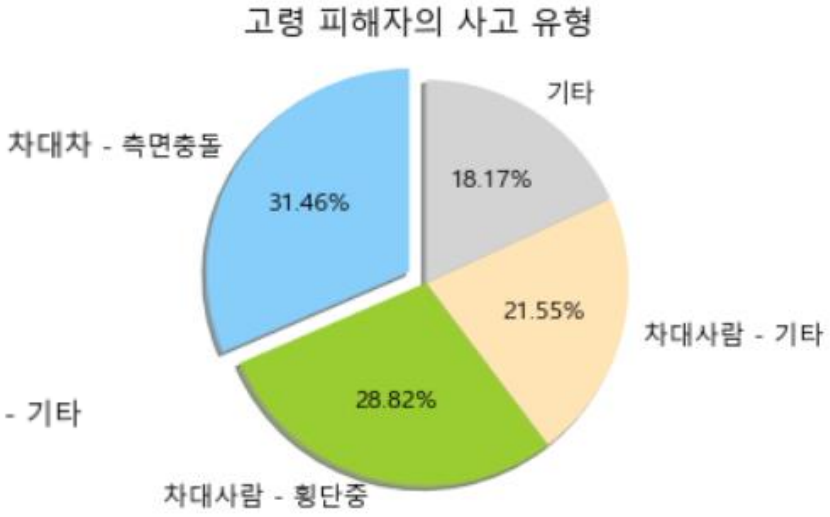
- 같은 사고 유형은 같은 색상으로 표시 / 상위 3개 사고 유형만 표시(나머지 기타)
- 유사한 사고유형 비율로 그룹화를 진행 후 분석하였음에도 불구하고 2~60대의 주요 사고 유형이 유사
- 0~10대와 고령(70대 이상)의 경우 다른 나이대와 달리 차대사람-횡단중의 사고 비율이 높음.
- 20~60대의 경우 다른 나이대와 달리 차대차-추돌,기타 의 사고비율이 높음

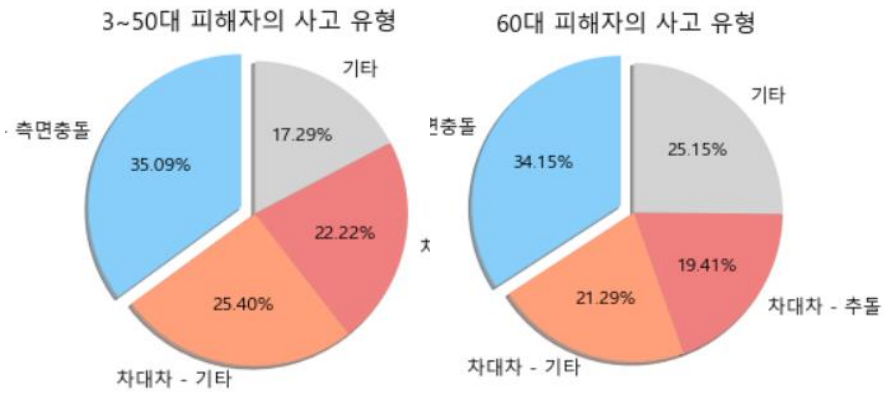
특정 나이대의 사고수를 줄이고 싶은 경우 아래의 사고 유형 감소에 집중

- 0~10대, 70~90대 차대사람-횡단중, 차대사람-기타, 차대차-측면충돌
- 20~60대 차대차





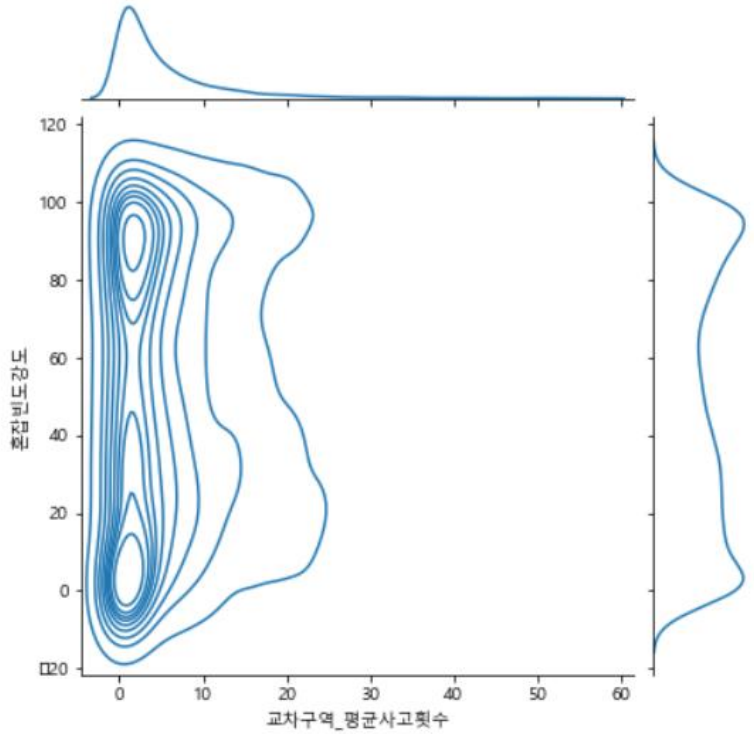




3. 도로의 혼잡빈도와 사고횟수 관계 분석

도로의 혼잡빈도강도와 해당 도로가 포함된 구역의 사고 횟수 사이의 관계 분석

- 그래프를 통해 도로의 혼잡빈도강도와 포함된 구역의 사고 횟수와는 관계가 없음을 확인할 수 있음.
- 추가로 피어슨, 스피어만 상관계수 검정을 통한 상관관계 분석을 진행하였으나, 각각의 상관계수가 0.04, 0.12 로 매우 작으므로 두 사이는 관계가 없음.

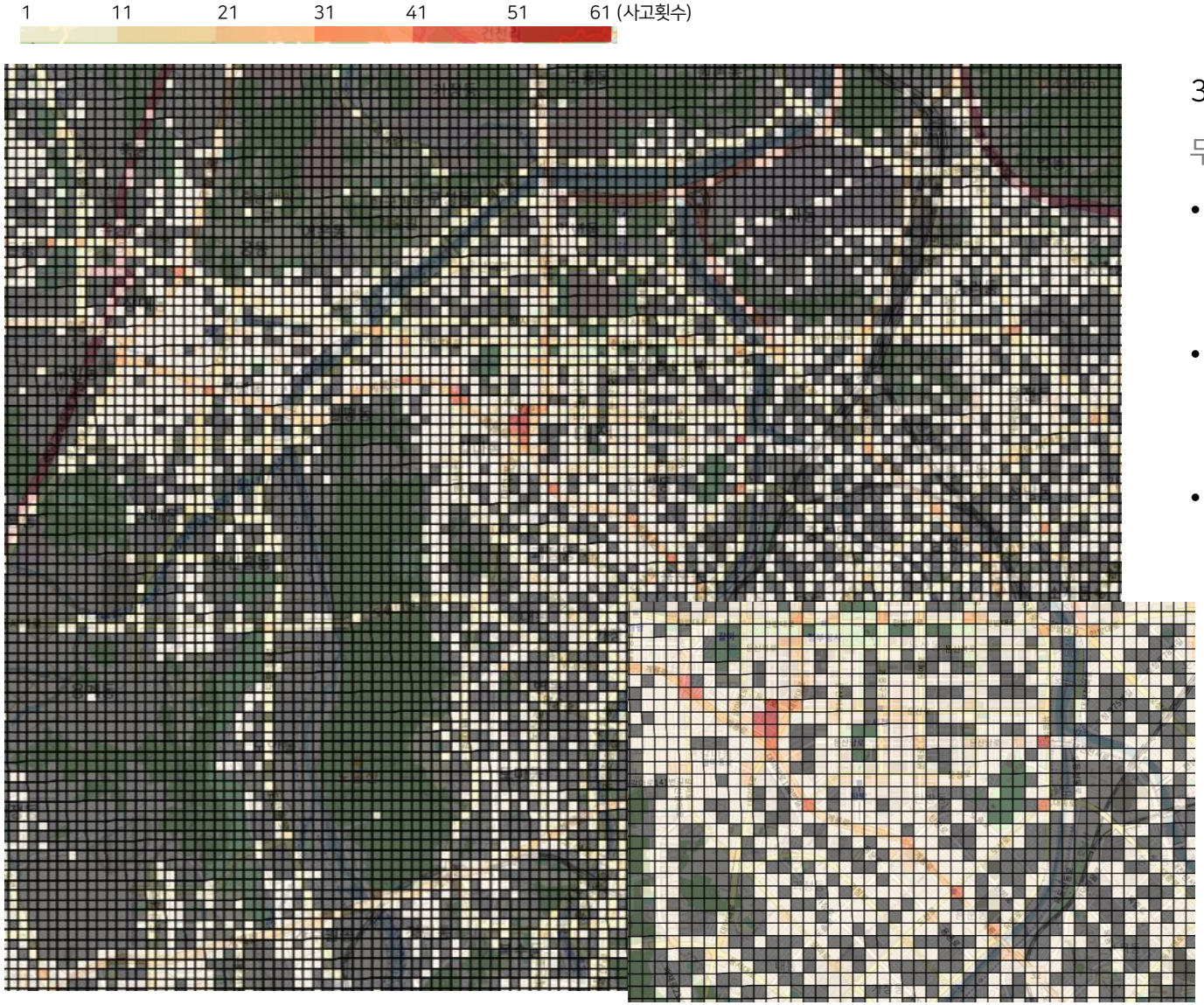


그래프1. 도로의 혼잡빈도강도와 해당 구역 평균사고횟수 의 Joint plot

4. 구역별 교통사고 횟수 분포 시각화

* 1. 1-2.ipynb 에서 진행

* 대전시 전체의 구역(GID)별 3년간 발생한 교통사고횟수를 인터랙티브하게 볼 수 있는 시각화파일('전체교통사고횟수시각화.html ')을 추가로 첨부 / 아래는 해당 파일의 일부



3년간 대전시 구역(gid)별 교통사고 발생횟수 분포를 시각화.

무사고 지역 - 사고적은지역 - 사고 많은 지역

- 전체 구역(54912개)에 대해 **3년간 사고가 한 번 이상 난 구역(**6068개) 은 **전체 구역의 11.5%**.

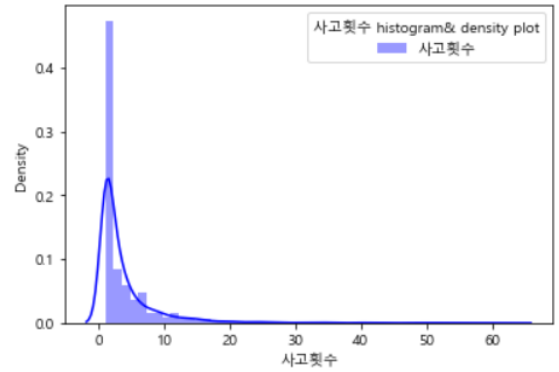
  -> 3년간 사고가 발생한 구역 비율이 매우 적음(11%).

- 사고가 1회 이상 발생한 지역에서 3회 이하 사고가 발생한 지역이 69%, 9회 이하 발생한 지역이 91% 차지

  -> 사고가 발생하더라도 대부분의 구역에서 사고가 적게 발생.

- 3년간 사고가 10회 이상 발생한 구역이 전체의 1%

  -> 매우 적은 특정 구역에서 사고가 많이 발생함.



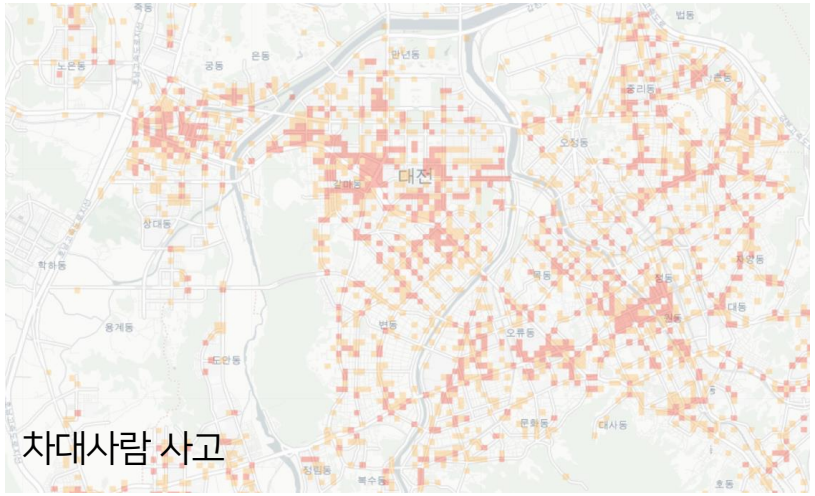
그래프1. 구역당 사고 밀도 그래프(사고가 1회 이상 일어난 지역 대상)

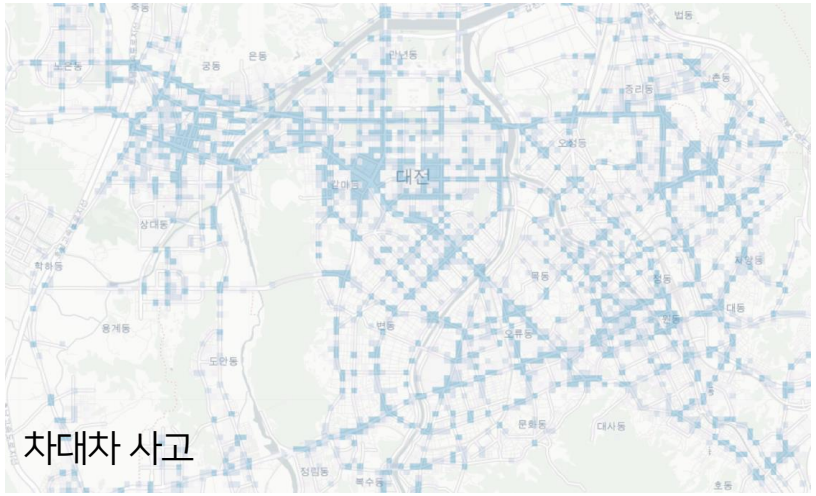
- 사고 발생횟수가 1-3회 사이에 매우 치우쳐져 있음.

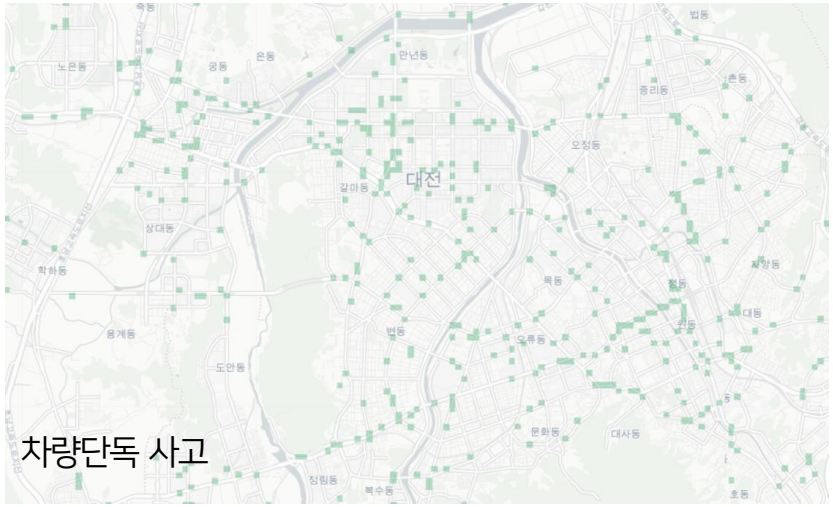
* 1-4..ipynb 에서 진행

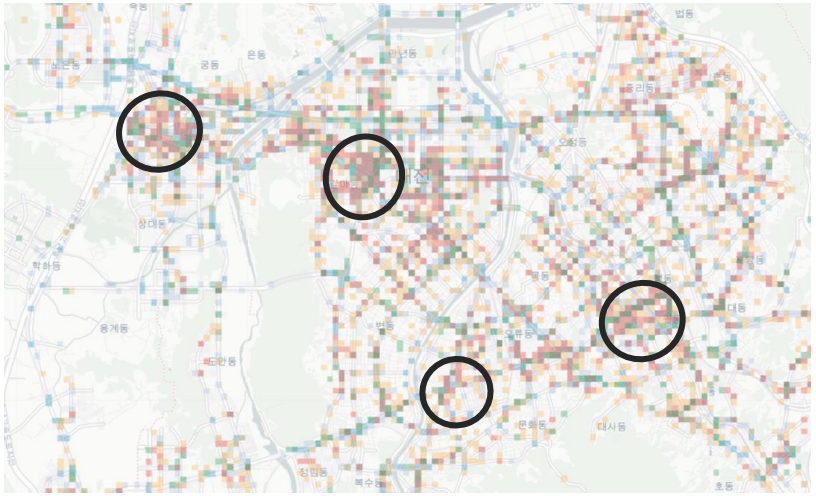
사고 유형별 사고횟수 시각화

- 사고 유형별 구역당 사고횟수를 시각화.
- 진할수록 사고 횟수가 높음(사분위수 별 색상 다르게 표시)



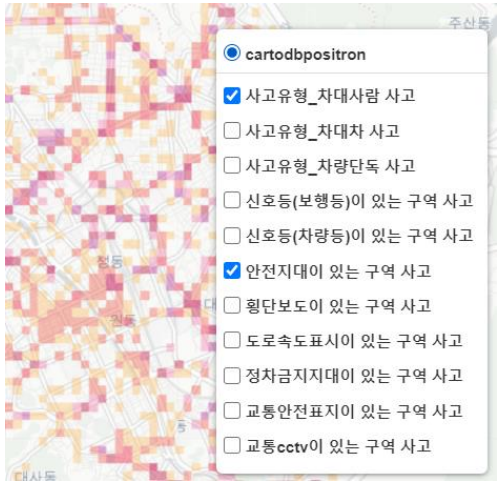






- 특정 사고 유형이 많이 발생했다고 해서, 다른 사고 유형도 많이 발생하는 것은 아님.
- 그러나, 사고유형에 관계 없이 사고가 많이 발생하는 몇가지 구역 존재.

- '/시각화자료/사고유형별.html ' 에서 인터랙티브하게 확인 가능.
- 우측의 레이어를 선택하여, 사고 유형별 사고 및 특정 교통안전물이 위치한 구역의 사고횟수도 확인 가능.



* 1-2.ipynb 에서 진행

5. 교통안전물-사고유형 인과관계 분석

* 교통안전물: 신호등(보행등/차량등), 안전지대, 횡단보도, 도로속도표시, 정차금지지대, 교통안전표지, 교통cctv, 중앙분리대

표1. (해당 교통안전물이 위치한 구역의 평균 사고횟수)/ (해당 교통안전물이 없는 구역의 평균사고횟수) 를 나타낸 표. 전체 구역 대상과 1회 이상 사고 구역 대상으로 분석.
- 수치가 높을수록 해당 교통안전물이 있는 구역의 평균 사고 횟수가 그렇지 않은 지역에 비해 높음.

	신호등(보행등)	신호등(차량등)	안전지대	횡단보도	도로속도표시	정차금지지대	교통안전표지	교통cctv	중앙분리대
(무사고지역포함) 사고횟수 비율	25.113952	26.436712	11.311592	41.015561	16.768968	15.808707	47.146061	38.516887	2.066940
(1회 이상 사고구역) 사고횟수 비율	1.869353	1.894579	1.290314	5.153348	1.598936	0.759344	7.839730	0.794234	0.744214

- 모든 구역 대상으로 분석 시, 모든 교통안전물에 대해 교통안전물이 있는 지역의 평균 교통사고 횟수가 그렇지 않은 지역에 비해 약 2~47배 높음.
  - 기존 연구에 따르면 교통안전물은 교통사고 감소 효과가 있으므로(상세 내용 appendix에 첨부), 이미 사고가 높게 나타나는 지역에 안전물을 설치했다고 가정
  - 사고가 발생한 구역이 약 12% 정도로 적으므로, 사고가 1회 이상 발생한 구역 대상으로 분석 재 진행.
- 1회 이상 사고 난 구역에 대해 분석 시,
  - 정차금지지대, 교통cctv, 중앙분리대가 설치된 구역은 그렇지 않은 구역에 비해 사고가 적게 발생.
    -> 정차금지지대, 교통cctv, 중앙분리대 는 사고 감소 효과가 큼 혹은 설치한 구역이 비교적 사고가 적게 나는 구역임
  - 다른 안전물의 경우 해당 안전물이 없는 구역보다 평균 교통사고 횟수가 1.2~1.9배 높게 발생.

데이터 분석의 한계점

1. 각 교통안전물의 설치 날짜를 데이터에서 확인할 수 없어 설치 전후 사고 데이터를 비교할 수 없고, 따라서 동일한 환경에서의 교통안전물의 효과를 비교하기 어려움
2. 단순히 데이터로 교통안전물과 사고유형 사이의 인과관계를 분석하기엔 다른 요인들과 복합적인 영향을 미칠 수 있음.

해결방안

다른 요인과의 복합적인 영향도 고려하는

DNN 모델을 학습시켜

이를 통해 인과관계 분석하고자 함

* 1-2.ipynb 에서 진행

5. 교통안전물-사고유형 인과관계 분석

* 교통안전물: 신호등(보행등/차량등), 안전지대, 횡단보도, 도로속도표시, 정차금지지대, 교통안전표지, 교통cctv, 중앙분리대

	신호등(보행등)	신호등(보행등)제외	신호등(차량등)	신호등(차량등)제외	안전지대	안전지대제외	도로속도표시	도로속도표시 제외
사고유형_차대사람 - 기타	0.0474(-17%)	0.057400	0.0459(-20%)	0.058200	0.0386(-47%)	0.073500	0.0527(-10%)	0.058700
사고유형_차대사람 - 길가장자리구역통행중	0.0049(-63%)	0.013400	0.0045(-67%)	0.013700	0.0038(-75%)	0.015300	0.0073(-40%)	0.012400
사고유형_차대사람 - 보도통행중	0.0081(+98%)	0.004100	0.008(+95%)	0.004100	0.0059(-15%)	0.007000	0.0081(+69%)	0.004800
사고유형_차대사람 - 차도통행중	0.0068(-43%)	0.012100	0.0071(-39%)	0.011800	0.0052(-63%)	0.014500	0.008(-32%)	0.011900
사고유형_차대사람 - 횡단중	0.0938(+112%)	0.044100	0.0912(+102%)	0.045100	0.0691(-11%)	0.077800	0.0848(+46%)	0.057800
사고유형_차대차 - 기타	0.2251(+86%)	0.120900	0.2251(+88%)	0.119300	0.2386(+41%)	0.168700	0.228(+63%)	0.139500
사고유형_차대차 - 정면충돌	0.0282(+149%)	0.011300	0.0287(+167%)	0.010700	0.0296(+70%)	0.017300	0.0278(+97%)	0.014100
사고유형_차대차 - 추돌	0.1988(+118%)	0.090900	0.2023(+132%)	0.087100	0.226(+79%)	0.125900	0.2008(+86%)	0.107700
사고유형_차대차 - 측면충돌	0.351(+133%)	0.150600	0.3507(+136%)	0.148200	0.3436(+44%)	0.237500	0.3443(+84%)	0.186200
사고유형_차대차 - 후진중충돌	0.0093(-21%)	0.011800	0.0088(-27%)	0.012100	0.009(-36%)	0.014300	0.0114(0%)	0.011400
사고유형_차량단독	0.0267(+45%)	0.018400	0.0276(+56%)	0.017600	0.0305(+32%)	0.023100	0.0267(+28%)	0.020800

표1. 특정 교통안전물이 위치한 구역과 없는 구역의 사고유형별 평균 사고횟수 비교.
- 5% 이상 사고감소효과가 나타나면 파란색으로 표기
- 1회 이상 사고 난 구역에 대해 분석
- 사고 감소효과가 나타나지 않은 교통안전물에 대한 수치는 제외.
- 모든 교통안전물에 대한 분석 수치는 **appendix**에 첨부

교통안전물 당 사고감소효과 있는 사고 유형 목록

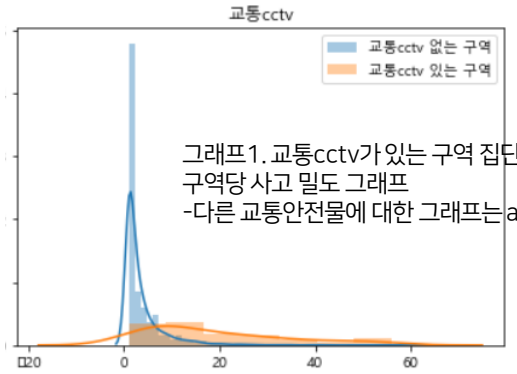
- **신호등**(보행등,차량등)
  차대사람- 차도통행중, 길가장자리구역통행중,
  차대차- 후진중충돌,
- **안전지대**
  차대사람-차도통행중, 보도통행중, 길가장자리구역통행중
  차대차-후진중충돌
- **도로속도표시**
  차대사람-길가장자리구역통행중,차도통행중

데이터의 한계와 교통안전물이 위치한 구역의 평균 사고 횟수가 약 12~1.9배가 **높은데도 불구하고**

특정 사고 유형이 적게 발생하는 신호등,안전지대,도로속도표시 교통안전물의 경우 **특정 유형의 사고 감소 효과가 있었다**고 판단.

T-test 통계 검정 시도

T-test: 두 집단을 비교하여 유의미한 차이가 있는 지 검정하는 통계 방법
- 각 교통안전물마다 교통안전물이 위치한 구역과 그렇지 않은 구역 집단 사이의 사고횟수에 유의미한 차이가 있는지 분석하는 t-test를 진행하고자 하였음.
- 하지만 t-test 의 선행조건인 정규성 검정에서, 각 집단이 정규성을 가지지 않아서 통계적 검정은 진행할 수 없었음.
- 정규성검정 방법과 결과는 appendix에 첨부.



그래프1. 교통cctv가 있는 구역 집단과 없는 구역 집단의 구역당 사고 밀도 그래프
-다른 교통안전물에 대한 그래프는 appendix에 첨부

위험지역의 정의

DNN 사고 예측 모델

위험 지역 ≠ 교통사고 많이 발생하는 구역

= 위험지수가 높은 구역(100m*100m)

= 학습된 모델에 의하면 사고가 많이 발생할 가능성이 높으나

현재까지는 많이 발생하지 않은 구역

위험 지수 = (DNN사고예측모델의 예측 사고 횟수) – (실제 교통사고 횟수)

- 위험구역 에 집중하면, 앞으로의 사고 발생 가능성을 줄일 수 있음.
- 이유: 현재 사고가 많이 발생하고 있는 다사고 구역이나, 특정 사고 유형, 특정 연령대의 사고가 많이 발생하는 구역은 현재 가지고 있는 데이터만으로 바로 조회가 가능하므로 탐색할 필요가 없음
- 위험 지수가 높으나 3년간 사고 횟수가 0인 구역은, 모델의 성능에 따른 오류라 판단하여 제외.

위험 구역 분석

1.위험 구역의 특징을 파악하기 위해 세가지 유형으로 군집화

2. 설명가능한 인공지능 기술(LIME)을 이용한 위험 구역별 분석

3. 교통안전물 증가에 따른 사고 감소 효과 분석

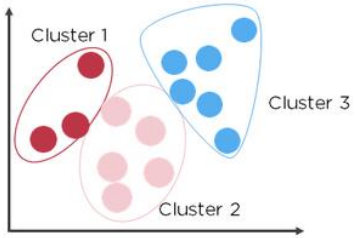
위험지역 군집화 분석

* 3-3.ipynb 에서 진행

상위 100개 위험지역

법정경계 (시군구)	위험구역 수
유성구	69
서구	50
대덕구	35
동구	28
중구	18

표1. 상위 100개 위험구역을 법정경계 (시군구) 로 분류.



위험지역을 다양한 관점에서 분석하기 위해 총 3가지 유형으로 **군집화**

1. 독립변수
2. 사고유형
3. 피해자 연령대

- 군집화 방법으로는 K-means Clustering 알고리즘을 채택
- 각 군집화마다 최적의 군집 개수를 찾기 위해 scree plot을 그려 탐색.

표2. 군집 유형별 사용 변수 종류.

군집 유형	사용 변수	변수 개수
독립변수	총인구, 교차로, 차량등록대수, 각 교통안전물 수 (횡단보도,교통cctv,신호등,도로속도표시, 교통안전 표지,정차금지지대,), 교통량, ic및 jc 개수,건물연면적, 속성변화점 수	11개
사고유형	차대차-추돌 /충돌/기타, 차대사람, 차량단독	5개
피해자 연령대	0-10대/ 20대/ 30-50대/ 60대/ 고령	5개

* 사용 변수는 앞선 데이터분석과 전처리 과정의 결과

위험지역 군집화 분석

* 3-3.ipynb 에서 진행

* 자세한 수치는 appendix에 첨부

1. 독립변수별 군집화

- 총 4가지 집단으로 분류 가능 (괄호 안은 구역당 집단별 평균 사고 횟수)

- **집단1. 도로구역** (36%)
  - 횡단보도 개수(4.6개)와 신호등 개수(5.9개),
  - 안전지대(1.6개) 와 교차로(0.9개)가 많음
  - 교통량이 중간 값이지만 교통안전표지 개수가 많음(10.8개)
  - 총 인구수(44.4명)와 건물연면적이 매우 적음(883m^2)

- **집단2. 교통량과 건물이 많은 구역** (31%)
  - 횡단보도 수가 적으나(3개), 교통량은 가장 많음
  - 차량등록대수가 가장 많음(165대), 건물연면적이 가장 큼.
  - 총인구수가 가장 많음.

- **집단3. 교통량 적은 구역** (26%)
  - 교통량이 제일 적으나 나머지 변수에 대해서는 중간 값 가짐

- **집단4. 거주지역** (6%)
  - 차량등록대수는 많으나(109대), 도로속도표시가 제일 적음(1.8개).

제안하는 위험지역의 경우,

36%가 도로구역으로 예측되며,

다음 31%는 교통량과 건물이 많은 지역이다.

26% 는 예상과는 달리 오히려 교통량이 적은 지역

* 3-3.ipynb 에서 진행

위험지역 군집화 분석

2. 사고 유형별 군집화

- 총 3가지 집단으로 분류 가능 (괄호 안은 구역당 유형별 평균 사고 횟수)

- **집단1: 차대차, 차량단독 사고 많은 지역.** (51%)
  - 차대차-추돌이 다른 집단에 비해 비교적 많이 발생(1.15건)
  - 동시에 차대차-충돌(평균3.18건) 많이 발생,
  - 차량 단독 사고가 다른 집단에 비해 많이 발생(0.28건).
  - 다만 차대 사람 유형의 사고는 비교적 적음.

- **집단2: 차대사람 사고 많은 지역.** (17%)
  - 차대사람 유형의 사고가 많이 발생(0.62건)
  - 비교적 다른 유형의 사고가 적음

- **집단3: 비교적 사고 적은 지역**. (31%)
  - 차대차-기타 유형이 많이 발생하는 곳(1.4건)
  - 모든 다른 유형의 사고가 적음

제안하는 위험지역의 경우,

51%가 차대차-차량단독 사고가 많이 일어났으며,

31%는 현재 차대차-기타 유형의 사고가 가장 많이 발생한다.

17%는 차대 사람의 사고가 많이 발생하는 지역으로 이 지역은 다른 유형의 사고가 적다.

* 3-3.ipynb 에서 진행

위험지역 군집화 분석

3. 피해자 연령대별 군집화

- 총 4가지 집단으로 분류 가능 (괄호 안은 구역당 집단별 평균 사고 횟수)

- **집단1: 생산가능인구 피해자가 가장 많은 구역.** (36%)
  - 30-50대 는 중간이지만 나머지 나이대의 거주자의 피해자가 가장 적음.
- **집단2: 고령화 지역.** (26%)
  - 20대 이하 피해자가 적은 편
  - 고령 피해자가 제일 많음(0.26명)
- **집단3. 교통사고 다발 구역** (6%)
  - 20대~60대 피해자가 가장 많음
  - 다른 나이대도 피해자가 많은 편
- **집단4. 10대 피해자 위험 구역** (31%)
  - 10대 이하 피해자 수가 가장 많음(0.27명)
  - 30-50대 피해자는 가장 적음(0.02명)

제안하는 위험지역의 경우,

36%는 생산가능 인구인 30-50대가 가장 사고를 많이 입은 지역,

31%는 10대 피해자가 가장 많은 구역

26%는 고령 피해자가 가장 많은 구역

위험지역 군집화 분석 결과

유사하게 분류된 군집

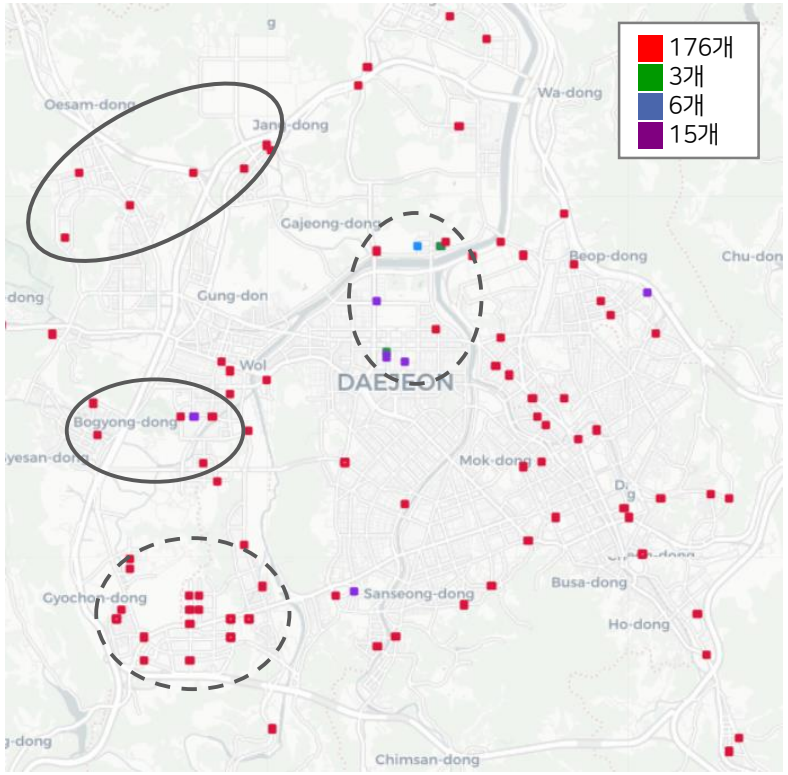
유형별 분류 차이가 큰 군집

* './시각화자료/위험지역_독립변수군집화/피해자연령대군집화/사고유형군집화.html' 에서 확인 가능.

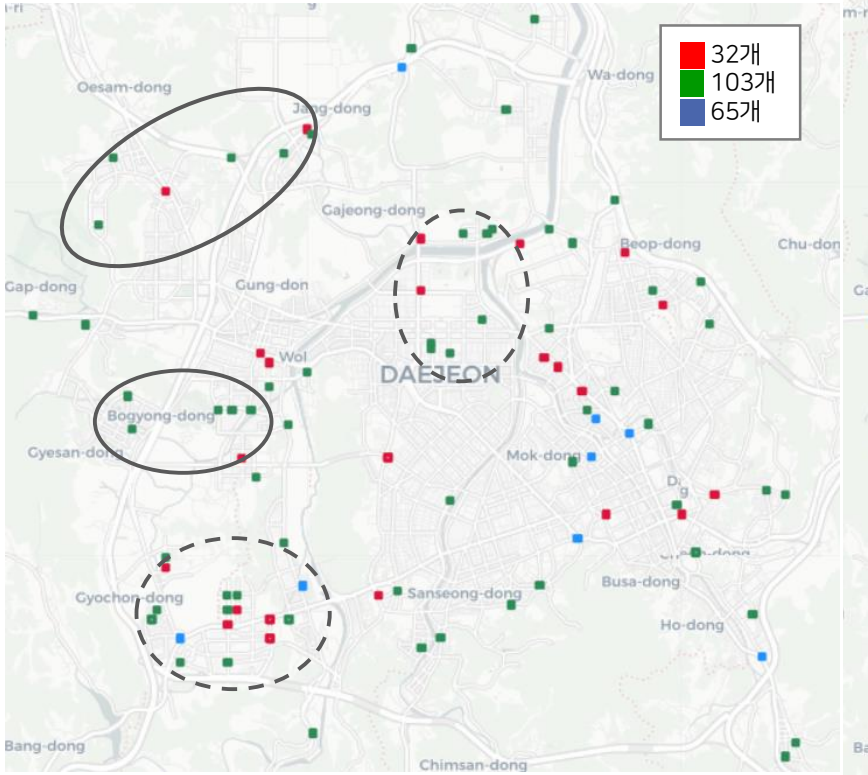
* 군집화 중심점에 대한 변수별 상세 수치는 appendix 에 첨부
* 다음은 좀 더 정확한 굽집화를 위해 상위 200개 구역에 대한 시각화 자료

* 3-3.ipynb 에서 진행

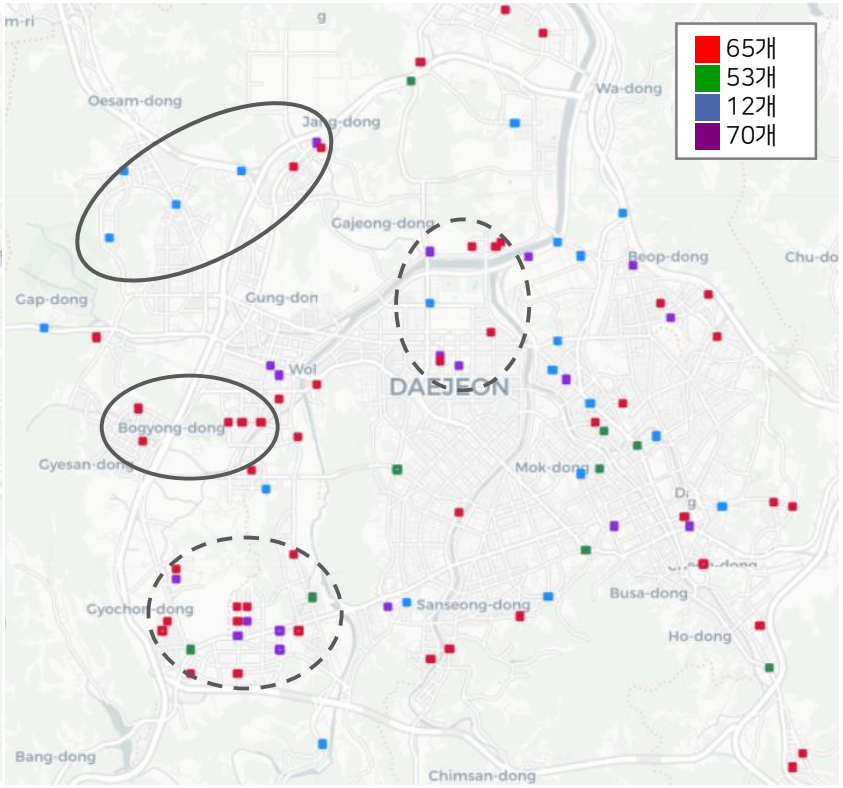
1. 독립변수별 군집화



2. 사고 유형별 군집화



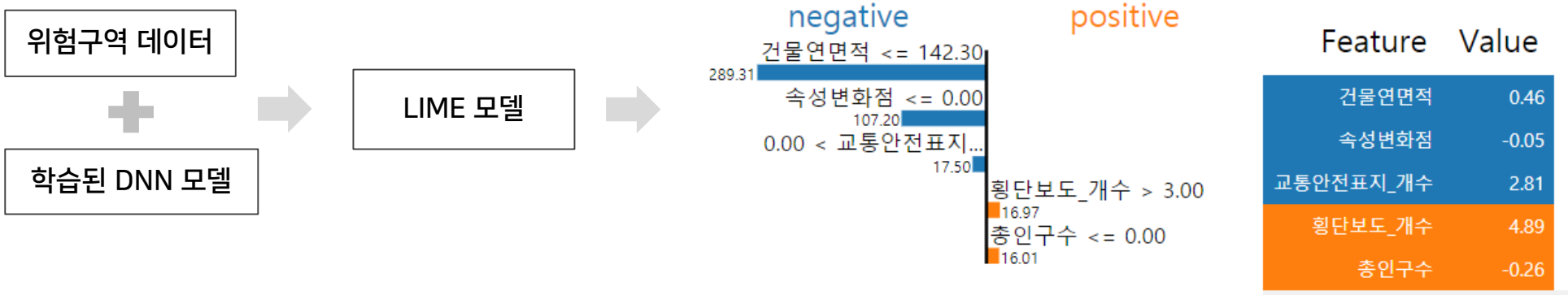
3. 피해자 연령대별 군집화



- 3가지 유형별로 군집화를 진행 - 각 유형별로 동일한 색은 동일한 군집에 있음을 의미
- 한 유형의 군집화에서는 동일한 군집에 있는 위험구역일지라도, 다른 유형별 군집화 결과는 다른 군집에 포함되어 있는 경우가 대부분.
- [결론] 상위 100개의 위험구역을 군집화하여 특징을 파악하는 것보다, 각 위험구역의 고유의 특징을 파악하는 것이 좋다.

  이 경우 군집화하는 경우 많은 특징이 누락되거나, 정확도를 높일 경우 너무 많은 군집이 생겨 군집화의 의미가 사라짐.

설명가능한 AI(XAI)- LIME 이용



- 위험구역을 군집화하여 특징을 파악하는 것에는 누락되는 정보가 많다고 판단하였고,
- 따라서 DNN 모델의 원인을 추론할 수 있는 XAI 기술을 이용해, 각 위험구역 별로 DNN사고예측 모델의 추론 과정을 파악하고자 하였다.
  - DNN 모델은 다른 요인과의 복합관계를 고려할 수 있어, 다른 모델에 비해 좋은 성능으로 사고를 예측할 수 있다.
  - 하지만 그 추론 원인을 파악하기 불가능 하다는 문제점이 있었다.
  - 최근 개발된 XAI 는 학습된 딥러닝 모델이 어떻게 추론했는지 분석 가능하게 하는 설명가능한 인공지능(Explainable AI, XAI)을 의미한다.
- LIME[23]은 이러한 XAI 기술 중 하나로, Python lime library를 이용하여 구현이 가능하다.

* 2-3.ipynb 에서 진행

위험지역에 영향을 미치는 요인 및 중요도

	중요도
중앙분리대_개수 <= 0.00	12.584078
총인구수 <= 0.00	8.505928
교통cctv_개수 <= 0.00	4.308191
속성변화점 <= 0.00	2.744334
0.00 < 총인구수 <= 4092.15	1.885607
교통cctv_개수 > 0.00	1.591884
교통안전표지_개수 > 14.16	1.057475
정차금지지대_개수 > 0.00	0.757003
안전지대_개수 <= 0.00	0.666678
1.24 < 횡단보도_개수 <= 3.17	0.571398
안전지대_개수 > 0.52	0.423656
도로속도표시_개수 > 0.74	0.305359
속성변화점 > 0.00	0.255051
횡단보도_개수 > 3.17	0.240694
0.00 < 신호등_개수 <= 2.05	0.159646
교차로 <= 0.00	0.022621

표1. 위험구역 가능성을 **증가**시키는 요인

* 위험구역일 가능성을 감소시키는 요인에 대한 상세 수치는 appendix에 첨부

- 각 구역마다 같은 요인일지라도 다른 요인들과 복합적인 작용을 일으켜 긍정적 혹은 부정적 영향을 미칠 수도 있음.
- 하지만 대략적으로 요인의 중요도를 알아보기 위해, 왼쪽 표는 상위 100개 위험지역의 사고횟수에 영향을 미친 요인의 평균 중요도 값을 나타낸 것.

위험구역일 가능성을 증가시키는 요인
(100m*100m구역당)

- 중앙분리대 없음
- 거주 인구 없음
- 속성변화점 없음
- 교통cctv 1개 이상
- 교통안전표지 14.16개 이상
- 정차금지지대 1개 이상
- 횡단보도 1.24개 이상
- 안전지대 없음
- 도로속도표시 0.74개 이상
- 신호등 2개 이하
- 교차로 없음

해석

현재 사고가 많이 일어나는 구역이 아닌

위험구역이 될 가능성이 높은 구역은

거주 인구가 없는 구역으로 도로가 있을 가능성이 큰 구역,

중앙분리대나 교차로가 있는 큰 도로보다는 작은 도로,

교통 cctv가 없어, 사고 인식이 적은 구역,

혹은 교통안전표지가 많은 구역

일 수 있다.

그러나 이 값은 위험 구역 요인의 평균 값이므로,

대략적인 값을 알기 위한 것이며

각 구역마다 다르게 영향을 미칠 수 있다.

- Lime의 parameter: num_samples = 10000, num_features=10

* 2-3.ipynb 에서 진행

[추가] 교통량이 0이 아닌 구역에 대한 위험지역 분석

	중요도
건물연면적 > 477.44	191.365391
cars_cnt > 12.51	1.276012
중앙분리대_개수 <= 0.29	0.999891
도로시설물 <= 0.01	0.626965
안전지대_개수 > 4.02	0.617548
정차금지지대_개수 <= 0.34	0.558759
중앙분리대_개수 > 0.29	0.419617
교통량4 <= 9.17	0.345106
횡단보도_개수 > 10.89	0.101797

표1. 위험구역 가능성을 **증가**시키는 요인

- 다음은 앞과 동일한 방법이지만, 교통량이 0인 구역을 제외하고 위험 지역을 구해 분석한 결과이다.
- 앞선 결과와 달리 건물연면적이 가장 영향력 절댓값이 큰 요인으로 나타남
  - 무사고 구역이 전체 구역의 89%를 차지하는데, 무사고 지역의 대부분이 산, 하천 등의 사람이 거주하거나 유동인구가 적은 지역일 것이라고 추측할 수 있음.
  - 산과 하천 등의 구역에는 교통량이 0이 아니어도 건물이 설치되어 있지 않을 가능성이 크고, 따라서 사고가 날 위험 지역일 가능성이 적다.
  - 반면, 건물이 많이 설치된 곳에는 유동인구와 교통량이 많으므로 사고가 날 가능성이 비교적 크므로 모델이 제대로 학습했다고 판단 가능.

위험구역일 가능성을 증가시키는 요인 (1개 구역당)

- 건물연면적 477.44㎡ 이상
- 자동차 등록대수가 12.51대 이상
- 중앙분리대 수 0.29개 이하
- 도로시설물 0.01개 이하
- 안전지대 개수가 4.02개 이상
- 정치금지지대 개수 0.34개 이하
- 교통량4 9.17 이하
- 횡단보도 개수 10.89 초과

교통량이 0이 아닌 구역에 대해

위험구역이 될 가능성이 높은 구역은

건물이 많고, 차량 수는 많지만,

여러 교통안전물 및 도로시설물 수는 적고,

도로 혼잡빈도는 적지만*,

횡단보도 개수는 많은 구역.

해석

유동인구와 차량은 많지만,

교통안전물이 적어

차량운전자 및 보행자의 사고

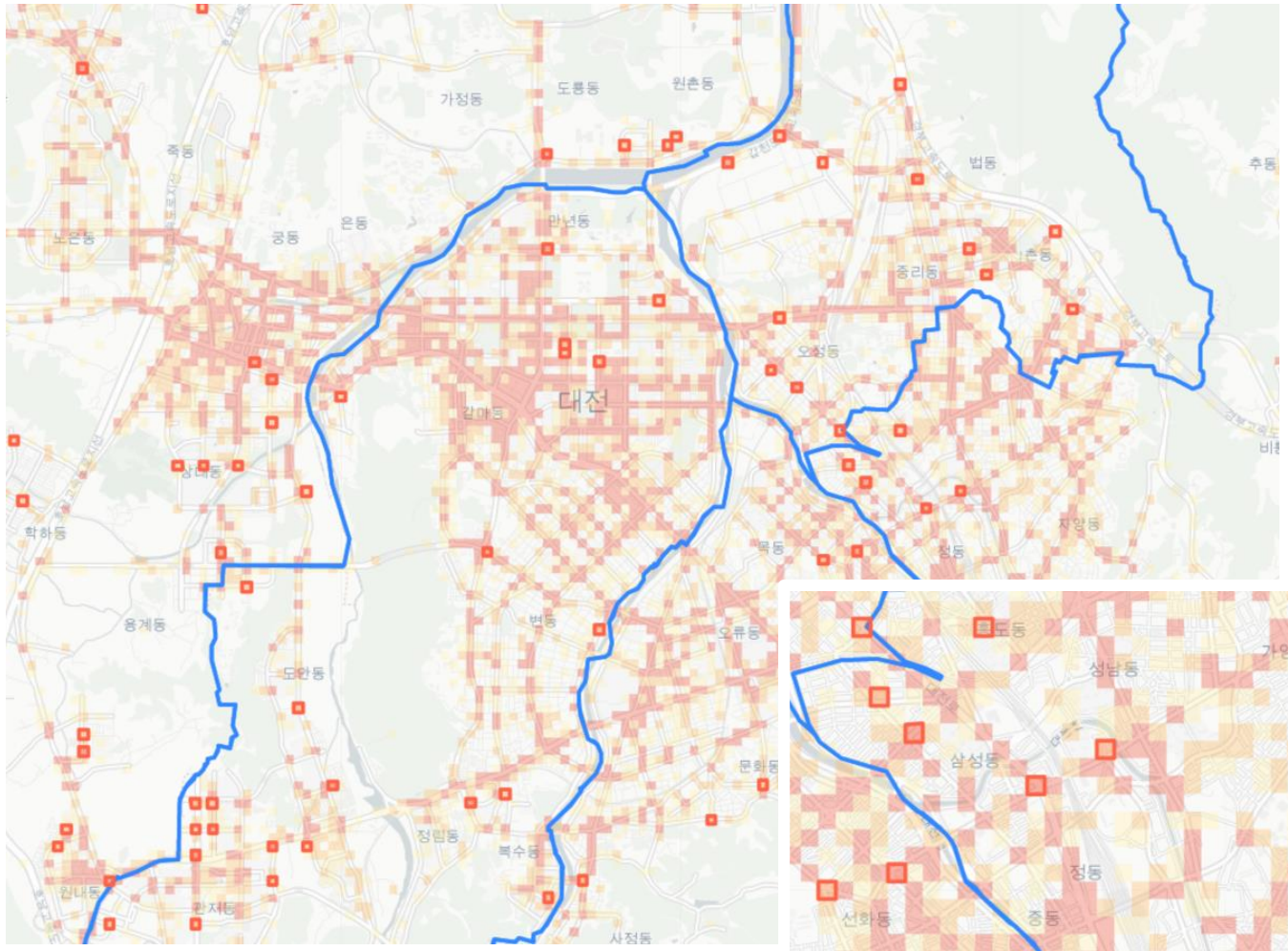
위험 인식이 적은 지역

- 교통량4 = 교차하는 모든 도로의 혼잡빈도 평균값 (자세한 내용 appendix 참고)
- 도로시설물 = 일반적으로 터널, 지하차도, 고가차도 를 의미하나 자세한 내용은 대전시 자료 참고

* 앞에서 진행한 데이터 분석 결과, 도로혼잡빈도와 교통사고 횟수는 관련이 없음.

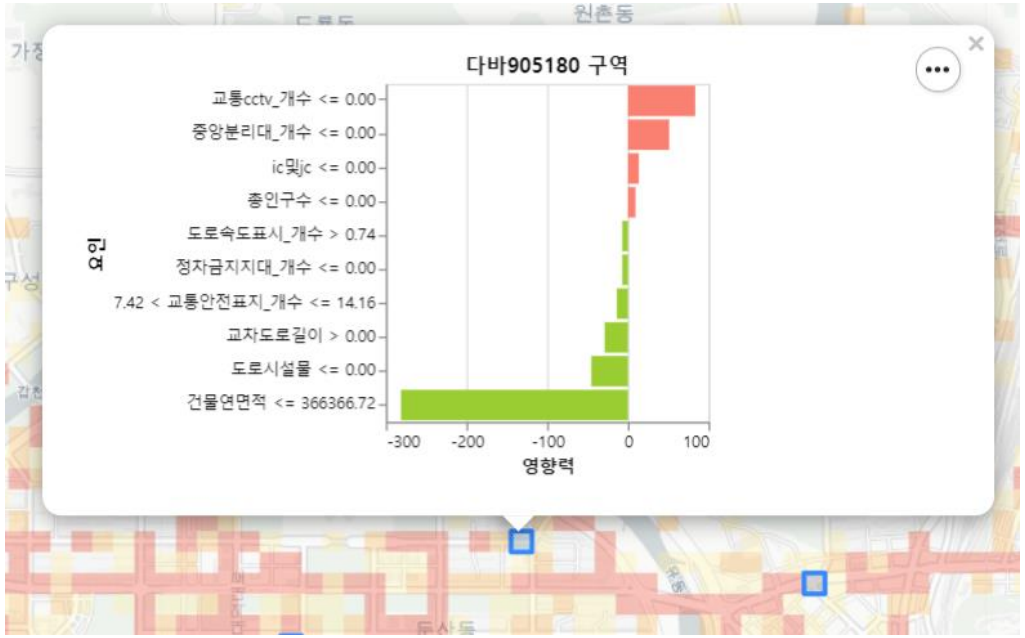
* 3-1.ipynb 에서 진행

위험지역에 영향을 미치는 요인 및 중요도



제안하는 위험 구역은 사고가 많이 발생한 구역(붉은 구역)과는 관련이 없음.

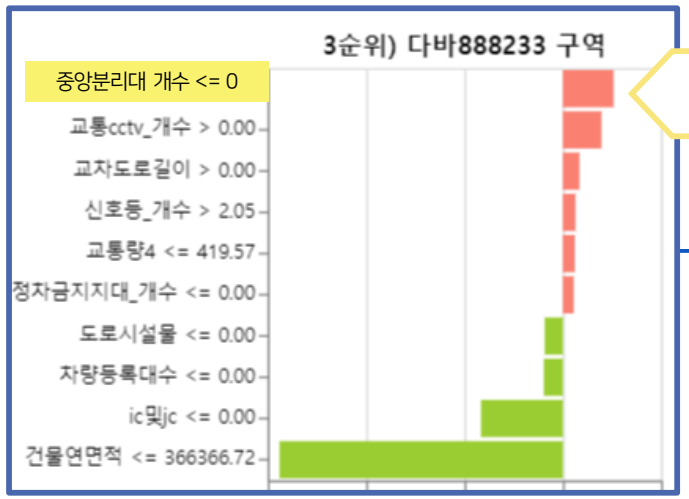
- 상위 100개 위험 구역(100m*100m)을 지도에 나타냄
- '시각화자료/위험지역_요인영향력그래프.html ' 에서 interactive web 확인 가능



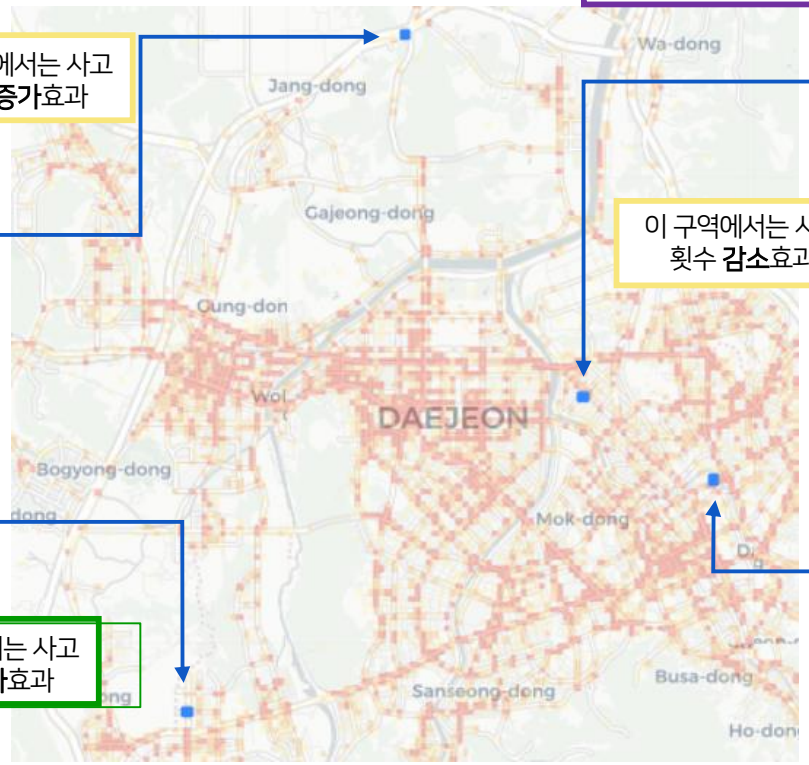
- 위험 구역을 **클릭**하면 아래처럼 해당 위험 구역에 영향을 미친 요인과 그 영향력 정도를 확인할 수 있다.
- 빨간 막대는 사고 횟수를 증가시키는 요인, 초록 막대는 감소시키는 요인이다.
- 막대가 길수록 영향력이 크다.

* 3-1.ipynb 에서 진행

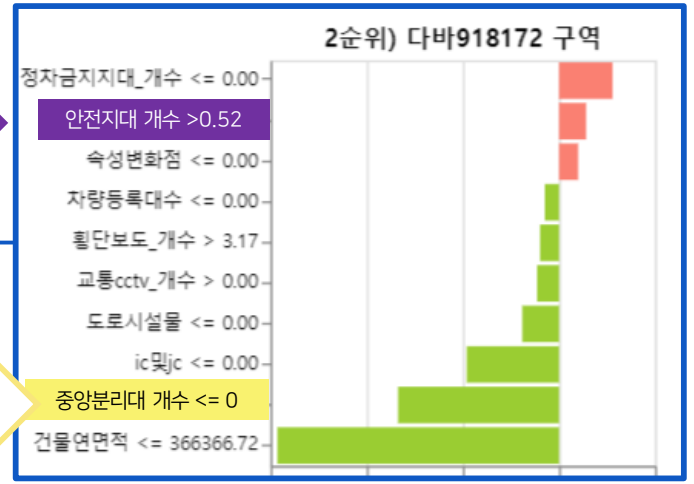
위험지역 상위 4개 구역 분석



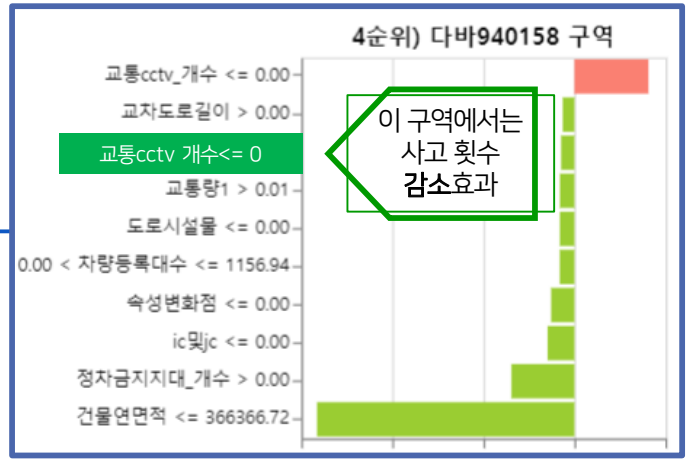
이 구역에서는 사고 횟수 **증가**효과



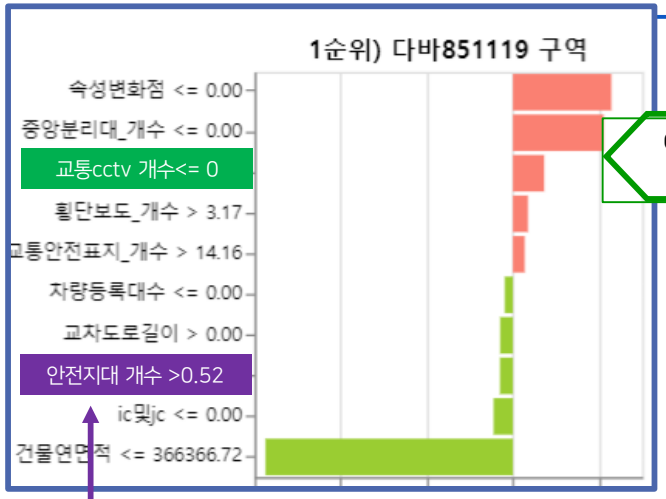
이 구역에서는 사고 횟수 **증가**효과



이 구역에서는 사고 횟수 **감소**효과



이 구역에서는 사고 횟수 **감소**효과



이 구역에서는 사고 횟수 **증가**효과

이 구역에서는 사고 횟수 **감소**효과

중앙분리대 개수 <= 0
교통cctv 개수<= 0
안전지대 개수 >0.52

예시로 세가지 요인을 보면 알 수 있듯이, 같은 요인일지라도 어떤 구역에는 사고 수를 높이는 요인으로 작용하지만, 다른 구역에는 오히려 사고 수를 감소시키는 요인으로 작용할 수 있다.

- 이는 학습이 안 된 것이 아니라, DNN모델이 다른 요인과의 복합적인 작용을 고려하여 제대로 학습된 것이다.
- 각 요인은 사고횟수와 무조건 양의 관계 혹은 음의 관계에 있다고 할 수 없으며. 정확한 분석을 위해선 다른 요인과의 관계를 살펴볼 필요가 있다.

* 3-2.ipynb 에서 진행

위험지역의 교통안전물 추가 효과 분석

구역별 특정 교통안전물 추가되었을 때의 교통사고 감소효과 정의

(구역별 교통안전물이 N개 추가 되었을 때의 효과) = (추가되지 않았을 때의 예측 교통사고 수) – (해당 교통안전물이 N개 추가되었을 때의 예측 교통사고 수)

- DNN 사고 수 예측 모델을 이용해 각 위험 지역에 특정 교통안전물이 N개 추가되었을 때의 예측 사고수를 알 수 있음.
- './effect_df2.pkl'에서 N=1~10 개 까지의 데이터를 확인할 수 있으며, 분석 결과에서는 각 교통안전물 1개를 추가했을 경우의 효과를 예로 보인다.
- 상위 100개 구역에서, 각 구역별로 교통안전물 추가 효과가 가장 큰 교통안전물과 그 효과를 *'./시각화자료/위험구역_추가교통안전물_효과.csv'* 에 첨부.

	gid	제안하는 추가 교통안전물	추가 교통안전물 사고감소효과
0	다바851119	횡단보도_개수+1	9.198707
1	다바918172	횡단보도_개수+1	8.485001
2	다바888233	횡단보도_개수+1	8.102486
3	다바940158	교통안전표지_개수+1	10.235548
4	다바969143	횡단보도_개수+1	7.172295
...	...	...	...
95	다바864158	횡단보도_개수+1	3.174795
96	다바851116	횡단보도_개수+1	2.702312
97	다바894174	횡단보도_개수+1	2.874895
98	다바896113	횡단보도_개수+1	3.253967
99	다바947139	횡단보도_개수+1	3.646323

100 rows × 3 columns

- 상위 100개 위험구역에 대해, 각 구역당 제안된 교통안전물은 다음과 같다.
  - 횡단보도 : 91개 구역
  - 교통안전표지: 7개 구역
  - 도로속도표시: 2개 구역
- 각 구역당 제안하는 교통안전물을 추가했을 경우, 3년간 최소1.41~최대10.8건의 평균 3.52건의 사고 감소 효과가 있다.
- 모델에 따르면, 아무 지역이나 단순히 교통안전물을 추가한다고 사고가 줄어드는 것이 아니라 위험 지역에 필요한 교통안전물을 추가해야 사고 감소 효과를 가질 수 있다.
  - 사고가 1회 이상 난 구역에 대해, 횡단보도와 도로속도표시 1개 추가 시 평균 1.6, 0.67회 사고 감소 효과를 보였으나 다른 안전물에 대해서는 오히려 평균사고 수가 증가.

    * appendix에 상세 수치 첨부
- 더불어 모델을 이용해 각 구역당 여러 교통안전물을 복합적으로 추가했을 때의 예측되는 사고 감소 효과도 확인할 수 있다.

추가 파일 목록

* 3-1.ipynb 에서 진행

- **위험지수가 높은 상위 100개 구역**(100m*100m, gid로 구별)과 각 위험지역의 **추가 설치 시 효과가 가장 좋은 교통안전물의 종류**와 해당 교통안전물의 사고 **감소효과**를 *'./시각화자료/위험구역_추가교통안전물_효과.csv'* 에서 확인 가능합니다.
- 앞선 ' 상위 4개 지역 분석' 과 같이 상위 100개 구역에 대한 분석 결과는 **interactive web 형태로 제공되며** ' 시각화자료/위험지역_요인영향력그래프.html ' 에서 조회할 수 있습니다..
- Folium library를 이용해 추가로 총 상위 200개 구역을 위험지수가 높은 순서대로 나열하여, pandas.DataFrame 형태로 *danger_200_df_gid.pkl'* 파일로 저장.
- 학습된 모델은 keras model load를 이용해 '512_256_128_9e_6_d09_batch32_3273'파일을 통해 불러올 수 있습니다. (test를 위해서는 코드와 동일한 방법으로 scaling이 필요합니다.)

4. Conclusion

결론

- 위험지역을 총 3가지 유형(독립변수, 사고유형, 피해자 연령대)으로 군집화 하였으며, 각 유형별로 3~4개 군집으로 나뉘어져 위험지역의 각 군집별 특징을 파악할 수 있다.
- 그러나 서로 다른 유형으로 군집화 하였을 경우 다른 군집에 포함되는 경우가 매우 많았음. 따라서 클러스터링 방법으로 위험지역의 특징을 한 번에 파악하기는 어렵다고 판단하였다.
- XAI 기술을 이용해 각 위험지역별 위험지역이라고 추론한 원인과 그 영향력을 확인할 수 있도록 데이터를 제공하였다.
- 전체적인 요인의 영향력을 보기 위해, 중요 요인과 그 평균 영향력을 제공한다.
- 각 위험지역 별로, 설치 시 효과가 큰 교통안전물 종류와 설치시 사고 감소 효과를 제공한다.

교통사고 예측 모델 활용법

- 교통사고 발생 건수를 예측하는 DNN 모델을 활용하여, 추가로 여러 분석을 진행할 수 있다.
  - 단순히 구역 당 한 개의 교통안전물의 추가 효과를 보는 것이 아니라, 여러 교통안전물을 동시에 추가하였을 때의 사고 감소 효과도 비교할 수 있다.
  - 총인구수와 건물 연면적에 변화를 주어 건물 설치 시, 예상 교통사고 증감 효과를 확인할 수 있다.
  - 도로 및 교차로 추가 설치시의 교통사고 증감 효과를 확인할 수 있다.
  - 같은 조건에서, 특정 변수 값을 다르게 하였을 때의 예상 사고 횟수를 확인할 수 있다.

의의

- 위험지역을 3가지 유형으로 군집화한 결과와 그 특징을 분석.
- XAI를 이용해 위험지역으로 각 위험지역별 추론하게 된 요인들과 그 영향력 제안
- (조사한 바에 따르면) 국내 데이터를 딥러닝 모델을 활용하여 넓은 범위의 교통사고를 예측한 최초의 연구.

감사합니다

APPENDIX

참고문헌

[3] J. Abellan, G. Lopez, AND J. D. Ona, Analysis of traffic accident severity using decision rules via decision trees, Expert Systems with Applications, 40(15) (2013), pp. 6047-6054.
[4] L. Lin, Q. Wang, AND A. W. Sadek, Data mining and complex network algorithms for traffic accident analysis, Transportation Research Record, 2460(1) (2014), pp. 128-136
[5] L.Y. Chang AND W.C. Chen, Data mining of tree-based models to analyze freeway accident frequency, Journal of safety research, 36(4) (2005), pp. 365-375.
[6] C. Caliendo, M. Guida, AND A. Parisi, A crash-prediction model for multilane roads, Accident Analysis and Prevention, 39(4) (2007), pp. 657-670.
[7] L. Lin, Q. Wang, AND A. W. Sadek, A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction, Transportation Research Part C: Emerging Technologies, 55 (2015), pp. 444-459.
[8]] L. Wenqi, L. Dongyu, AND Y. Menghua, A model of traffic accident prediction based on convolutional neural network, 2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE), (2017), pp. 198-202.
[9] J. D. Ona, R. O. Mujalli, AND F. J. Calvo, Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks, Accident Analysis and Prevention, 43(1) (2011), pp. 402-411.
[10] A. Iranitalab AND A. Khattak, Comparison of four statistical and machine learning methods for crash severity prediction, Accident Analysis and Prevention, 108 (2017), pp. 27-36.
[11] H. Ren, Y. Song, J. Wang, Y. Hu, AND J. Lei, A deep learning approach to the citywide traffic accident risk prediction, 2018 21st International Conference on Intelligent Transportation Systems (ITSC), (2018), pp. 3346-3351.
[12] Z. Yuan, X. Zhou, AND T. Yang, Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data, Proceedings of 38 the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2018), pp. 984-992.
[13] Z. Zhou, Attention Based Stack ResNet for Citywide Traffic Accident Prediction, 2019 20th IEEE International Conference on Mobile Data Management (MDM), (2019), pp. 369-370.
[14] J. Bao, P. Liu, AND S. V Ukkusuri, A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data, Accident Analysis and Prevention, 122 (2019), pp. 239-254.
[15] 이현미(Hyun-Mi Lee), 전교석(Gyo-Seok Jeon) , 장정아(Jeong-Ah Jang). "LightGBM 알고리즘을 활용한 고속도로 교통사고심각도 예측모델 구축." 한국전자통신학회 논문지, (2020): 1123-1130
[16] 김호용, et al. "교통사고 지점 예측을 위한 인공신경망 모델." 한국지능정보시스템학회 학술대회논문집 (2017): 33-34.
[17] Doorham Bae, Byeongchan Seong. (2019). Multiple aggregation prediction algorithm applied to traffic accident counts. The Korean Journal of Applied Statistics, 32(6), 851-865.
[18] 홍지연, 이수범, 김정현. (2015). 소규모 지역단위 교통사고예측모형 개발 - 서울시 행정동을 대상으로. 대한토목학회논문집, 35(6), 1297-1308.
[19] Wier M, Weintraub J, Humphreys EH, Seto E, Bhatia R. An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. Accid Anal Prev. 2009 Jan;41(1):137-45. doi: 10.1016/j.aap.2008.10.001. Epub 2008 Nov 4. PMID: 19114148.
[20] 류종득, et al. "딥 러닝을 이용한 고속도로 교통사고 건수 예측모형 개발에 관한 연구." 한국 ITS 학회 논문지 17.4 (2018): 14-25.
[21] 이홍석, 이주영. (2017). 와이드-앤-딥러닝 모델을 이용한 보행자 사고 심각도 예측. 한국정보과학회 학술발표논문집, (), 731-733.
[22] 김현용. (2020). GRU 기반 지역별 교통사고 심도 예측 모형 (Doctoral dissertation, 서울대학교 대학원).
[23] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.
[24] 이동민, 김도훈, and 송기섭. "비교그룹방법을 이용한 교통안전 시설물 설치 효과 분석." 대한교통학회지 29.3 (2011): 31-40.
[25] 홍지연, 이수범, and 김정현. "소규모 지역단위 교통사고예측모형 개발-서울시 행정동을 대상으로." 대한토목학회논문집 35.6 (2015): 1297-1308.
[26] Wier, Megan, et al. "An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning." Accident Analysis & Prevention 41.1 (2009): 137-145.

1. 교통량 변수 계산 방법

[1_3.교통분석.ipynb 의 5번째 셀]

- 여러 교통량 데이터 중 '21.대전광역시_평일_일별_혼잡빈도강도(2018).csv ' 의 '일별 혼잡 빈도 강도'의 1년 평균치 를 채택
- 상행, 하행으로 나눠지는 도로의 경우 평균값 채택.

[2_1. gid별로_데이터통합.ipynb 의 9번째 셀]

- 교통량은 정의에 따라 달라질 수 있으므로 다음과 같이 네 가지 유형으로 정의 후 계산

1. [정의1] 구역당 교통량 = (해당 구역과 교차하는 도로 길이) * (해당 도로의 혼잡빈도) 의 총합
2. [정의2] 구역당 교통량 = (해당 구역과 교차하는 도로 길이) * (해당 도로의 혼잡빈도) 의 평균
3. [정의3] 구역당 교통량 = 교차하는 모든 도로의 혼잡빈도 최대값
4. [정의4] 구역당 교통량 = 교차하는 모든 도로의 혼잡빈도 평균값

- 상관분석 결과 교통량2,3,4가 서로 0.77 이상의 높은 상관계수를 가지므로 교통량 2,3 변수 제거 -> 교통량 1,교통량4 만 채택

2. 피해자 연령대 - 사고 비율 분석

피해운전자 연령대 / 사고유형	0~10대	20대	3~50대	60대	고령
차대사람 - 기타	18.993352	8.867261	5.154410	8.327933	20.710974
차대사람 - 길가장자리구역통행중	4.178538	2.946742	0.807228	0.874919	nan
차대사람 - 보도통행중	2.943970	0.729927	0.585055	0.810110	nan
차대사람 - 차도통행중	3.038936	1.946472	0.903503	1.166559	4.868624
차대사람 - 횡단중	35.232669	7.650716	5.628379	10.661050	27.279753
차대차 - 기타	14.434948	22.708840	25.505443	21.451717	17.156105
차대차 - 정면충돌	1.709402	2.595296	2.732726	2.948801	nan
차대차 - 추돌	nan	17.383077	22.069170	19.215813	nan
차대차 - 측면충돌	19.468186	32.603406	34.940384	33.895010	29.984544
차대차 - 후진중충돌	nan	2.568262	1.673702	0.648088	nan

표1. 피해자 연령대 그룹에 따른 사고 유형 비율 분석

- 각 사고 유형의 해당 피해자 연령대의 비율(%)
- 차량단독의 경우 피해자가 없기 때문에 제거

표2. 피해자 연령대 비율이 비슷한 사고 유형 상관 분석.

- 피해자 연령대 비율이 비슷할 수록 수치가 높음
- 차대사람의 경우 세부 사고 내용과 상관 없이 피해자 나이대 비율이 매우 유사.
- 차대차의 경우 후진중충돌을 제외하고 모든 사고의 나이대 비율이 매우 유사.

사고유형	차대사람 - 기타	차대사람 - 길가장자리 구역통행중	차대사람 - 보도 통행중	차대사람 - 차도 통행중	차대사람 - 횡단중	차대차 - 기타	차대차 - 정면충돌	차대차 - 추돌	차대차 - 측면충돌	차대차 - 후진중충돌
차대사람 - 기타	1.000000	0.880382	0.980765	0.969922	0.928431	-0.857006	-0.914139	-0.971943	-0.631681	0.113951
차대사람 - 길가장자리 구역통행중	0.880382	1.000000	0.812730	0.969291	0.788915	-0.801672	-0.891792	-0.798076	-0.864576	0.840574
차대사람 - 보도통행중	0.980765	0.812730	1.000000	0.924020	0.998613	-0.965800	-0.935971	-0.701421	-0.994593	-0.427537
차대사람 - 차도통행중	0.969922	0.969291	0.924020	1.000000	0.910845	-0.922793	-0.928558	-0.936373	-0.952277	0.652050
차대사람 - 횡단중	0.928431	0.788915	0.998613	0.910845	1.000000	-0.960353	-0.916312	-0.490110	-0.852330	-0.649511
차대차 - 기타	-0.857006	-0.801672	-0.965800	-0.922793	-0.960353	1.000000	0.840644	0.763622	0.894690	0.343135
차대차 - 정면충돌	-0.914139	-0.891792	-0.935971	-0.928558	-0.916312	0.840644	1.000000	0.125389	0.963003	-0.974588
차대차 - 추돌	-0.971943	-0.798076	-0.701421	-0.936373	-0.490110	0.763622	0.125389	1.000000	0.984026	-0.344438
차대차 - 측면충돌	-0.631681	-0.864576	-0.994593	-0.952277	-0.852330	0.894690	0.963003	0.984026	1.000000	-0.506068
차대차 - 후진중충돌	0.113951	0.840574	-0.427537	0.652050	-0.649511	0.343135	-0.974588	-0.344438	-0.506068	1.000000

국내 교통안전물의 효과 연구 조사

1.비교그룹방법을 이용한 교통안전 시설물 설치 효과 분석(2011) [24]

- 비교그룹 방법을 적용하여 9가지 교통안전 시설물에 대한 효과분석
  - 미끄럼방지시설: 2차로도로, 곡선반경있음, 종단경사있는 구역에 대해 89% 사고절감효과(곡선반경없을시 28% 절감)
  - 보행자보호시설: 모두 사고 증가, 다만 곡선이 없는 도로는 곡선 있는 도로보다 효과적. 2차로보다 4차로에서 효과적
  - 부가차로: 사고절감효과 보임, 특히 평면곡선부에서 효과적.
  - 중앙분리대: 평면곡선부에서 사고 감소에 특히 효과적, 곡선도로에 설치시 사고심각도는 떨어지나 사고 횟수는 증가
  - 버스정차대 위치조정과 우회전전용차로 설치는 사고 감소 효과 보임.

〈표 1〉 사고유형에 따른 교통안전 시설물 효과도

시설명	사고유형		사고변화율(%)
중앙분리대	정면충돌	직선·곡선	-80 ~ -70
	후미추돌	곡선	-5 ~ 20
가드레일	후미추돌	곡선	-85 ~ -55
	측면추돌	직선·곡선	-40 ~ -15
미끄럼방지포장	후미추돌	직선·곡선	-90 ~ -60
갈매기표지	후미추돌	곡선	-90 ~ -40
	측면충돌	직선·곡선	-60 ~ -45
속도규제표지	후미추돌	직선·곡선	-15 ~ 30

주 : 삼성교통안전문화연구소, "도로안전시설물의 사고감소 효과도 분석", 2004

2. 소규모 지역단위 교통사고예측모형 개발 - 서울시 행정동을 대상으로(2015) [25]

- 연구 결과에 따르면 교통안전물을 교통사고 발생횟수와의 상관관계로 분류 가능.
  - 사고와 양의 관계: 도로연장, 건축물 총 연면적, 버스전용차로 설치율, 교차로 및 횡단보도 개수
  - 사고와 음의 관계: 횡단보도예고 설치율, 과속방지턱 개소수

〈표 2〉 국내외 교통안전 시설물 효과분석 결과

분석방법	연구자	개선시설물	개선 효과분석
단순사고건수비교	김경석, 강승림 (2003)	중앙분리대	총교통사고 -35.5% 주간사고 -31.3% 야간사고 -43.2% 인피사고 -36.2% 정면충돌 -53.5% 추돌 -34.9% 차량단독 -18.7%
	이수범, 박규영 (2000)	중앙분리대	총사고건수 -48.20%
		노면요철포장	총사고건수 -42.1%
		안전카메라	신호감시 -25.0% 속도감시 -28.0%
		가드레일	총사고건수 -31.03
비교그룹	정도영 (2008)	어린이보호구역	총사고건수 -29%
	이동민 등 (2007)	노면요철포장	차도이탈건수 -38%
	Griffith (1997)	노면요철포장	전체이탈사고 -18.3% 부상이탈사고 -13% 지방부전체이탈사고 -21%
경험적베이즈법	정국영 (2008)	노면요철포장	총사고건수 -14% 졸음사고 -28%
	박규영 (2006)	중앙분리대	총사고건수 -18.46%
		가드레일	총사고건수 -38.11%
		갈매기표지	총사고건수 -16.24%
	Hirst 등 (2005)	안전카메라	1mph 속도감소에 총사고건수 -4% 감소
	Persaud 등(2004)	신호등	4지: 직각 -67% 추돌 138% 3지: 직각 -34% 추돌 150%
		노면요철포장	부상사고 -14% 정면/대향측면 -25%
	Elvik (1997)	안전카메라	부상사고 -20% 물피사고 -12%

교통안전물-사고횟수 분석

	신호등(보행등)	신호등(보행등)제외	신호등(차량등)	신호등(차량등)제외	안전지대	안전지대제외	횡단보도	횡단보도제외
사고유형_차대사람 - 기타	0.047417	0.057423	0.045949	0.058152	0.038632	0.073453	0.065640	0.021887
사고유형_차대사람 - 길가장자리구역통행중	0.004855	0.013436	0.004497	0.013668	0.003761	0.015316	0.009854	0.006287
사고유형_차대사람 - 보도통행중	0.008118	0.004087	0.008020	0.004095	0.005926	0.007033	0.008263	0.001211
사고유형_차대사람 - 차도통행중	0.006752	0.012057	0.007121	0.011774	0.005242	0.014534	0.011291	0.004890
사고유형_차대사람 - 횡단중	0.093771	0.044089	0.091223	0.045150	0.069060	0.077777	0.096125	0.010524
사고유형_차대차 - 기타	0.225097	0.120875	0.225096	0.119273	0.238632	0.168733	0.225558	0.043681
사고유형_차대차 - 정면충돌	0.028222	0.011290	0.028708	0.010750	0.029630	0.017347	0.025609	0.004377
사고유형_차대차 - 추돌	0.198847	0.090886	0.202309	0.087074	0.225983	0.125912	0.183269	0.038605
사고유형_차대차 - 측면충돌	0.350960	0.150608	0.350723	0.148196	0.343590	0.237497	0.333795	0.049828
사고유형_차대차 - 후진중충돌	0.009256	0.011801	0.008770	0.012081	0.009003	0.014274	0.013446	0.004238
사고유형_차량단독	0.026705	0.018392	0.027584	0.017609	0.030541	0.023130	0.027149	0.008522

	도로속도표시	도로속도표시제외	정차금지지대	정차금지지대제외	교통안전표지	교통안전표지제외	교통cctv	교통cctv제외	중앙분리대	중앙분리대제외
사고유형_차대사람 - 기타	0.052747	0.058652	0.100467	0.096739	0.063826	0.018993	0.027504	0.096834	0.000000	0.099443
사고유형_차대사람 - 길가장자리구역통행중	0.007348	0.012437	0.051402	0.017295	0.010030	0.005349	0.002821	0.018291	0.000000	0.018592
사고유형_차대사람 - 보도통행중	0.008135	0.004811	0.014019	0.010264	0.008014	0.000907	0.002821	0.010363	0.000000	0.010632
사고유형_차대사람 - 차도통행중	0.008048	0.011925	0.023364	0.017862	0.011613	0.003762	0.004231	0.018064	0.000000	0.018479
사고유형_차대사람 - 횡단중	0.084762	0.057833	0.084112	0.116983	0.090412	0.009746	0.049365	0.114899	0.052632	0.119286
사고유형_차대차 - 기타	0.228044	0.139516	0.212617	0.297250	0.227757	0.026608	0.221439	0.284218	0.210526	0.302991
사고유형_차대차 - 정면충돌	0.027817	0.014074	0.028037	0.032946	0.025434	0.002856	0.041608	0.030240	0.000000	0.033716
사고유형_차대차 - 추돌	0.200752	0.107733	0.135514	0.246215	0.192389	0.017724	0.228491	0.230817	0.157895	0.250000
사고유형_차대차 - 측면충돌	0.344297	0.186192	0.329439	0.421491	0.328582	0.032954	0.384344	0.398041	0.578947	0.430009
사고유형_차대차 - 후진중충돌	0.011372	0.011413	0.011682	0.019733	0.012765	0.003944	0.008463	0.019310	0.000000	0.020071
사고유형_차량단독	0.026679	0.020830	0.009346	0.040147	0.029177	0.004714	0.028914	0.037998	0.000000	0.040482

- 1회 이상 사고 난 구역에 대해
- 각 교통안전물이 위치한 구역과 해당 교통안전물이 없는 지역의 교통사고유형별 평균 사고 횟수를 나타냄.
- 어떤 교통안전물이 위치한 구역의 특정 사고유형이 해당 교통안전물이 없는 구역에 비해 5% 이상 감소하면 파란색으로 강조.

교통안전물 위치한 구역 수치분석 및 정규성 검정

<사고가 한 번 이상 발생한 모든 gid에 대해서만 각 교통안전물들이 위치한 gid 개수 및 비율>

- 신호등(보행등) --> 1990/6068 = 32.7%
- 신호등(차량등) --> 1951/6068 = 32.1%
- 안전지대 --> 1612/6068 = 26.5%
- 횡단보도 --> 3890/6068 = 64.1%
- 도로속도표시 --> 1955/6068 = 32.2%
- 정차금지지대 --> 51/6068 = 0.8%
- 교통안전표지 --> 4477/6068 = 73.7%
- 교통cctv --> 75/6068 = 1.2%

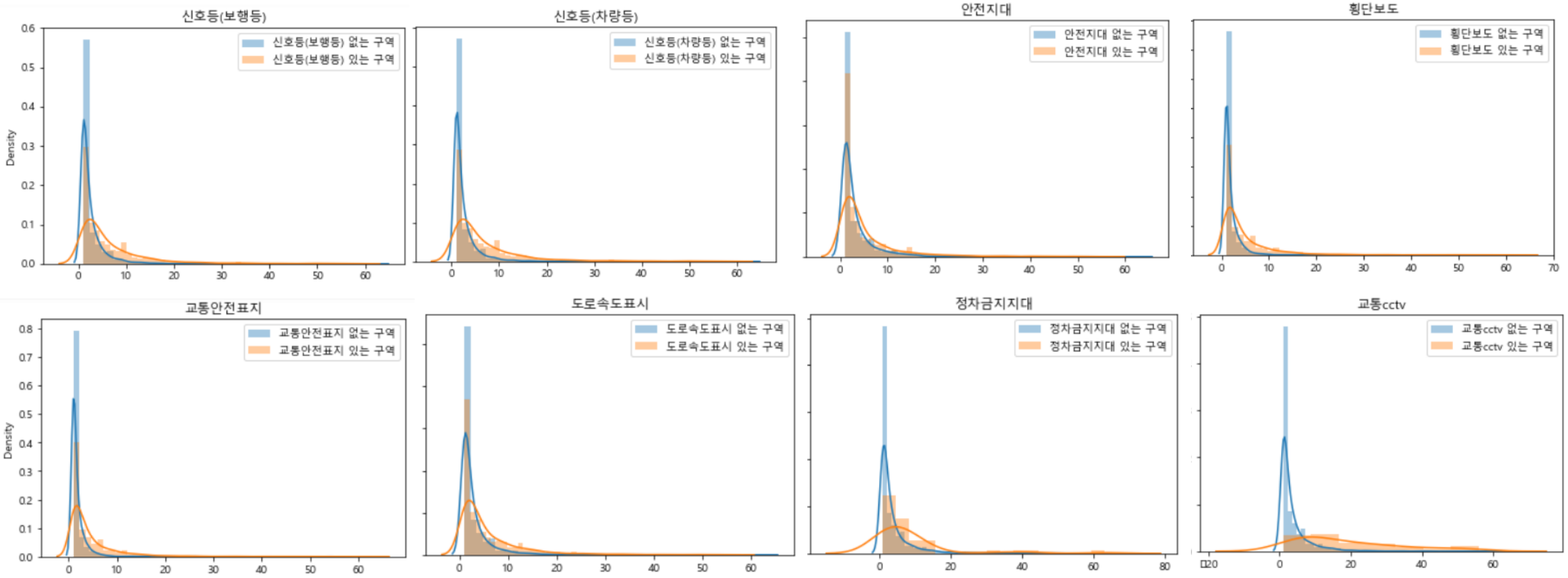
<모든 gid에 대해 각 교통안전물들이 위치한 gid 개수 및 비율 분석>

- 신호등(보행등) --> 2548/54912 = 4.6%
- 신호등(차량등) --> 2486/54912 = 4.5%
- 안전지대 --> 2631/54912 = 4.7%
- 횡단보도 --> 6012/54912 = 10.9%
- 도로속도표시 --> 2834/54912 = 5.1%
- 정차금지지대 --> 57/54912 = 0.1%
- 교통안전표지 --> 8375/54912 = 15.2%
- 교통cctv --> 81/54912 = 0.1%

교통안전물 유무 집단 정규성 검정

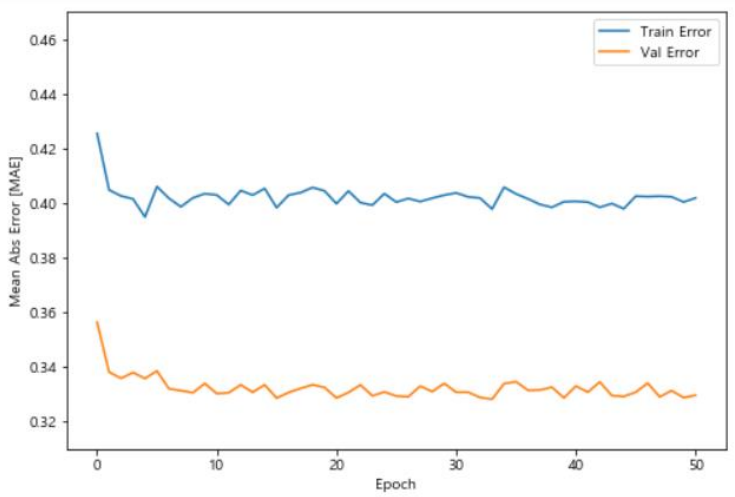
- 각 교통안전물이 위치한 구역과 아닌 구역의 차이가 있는지 확인하는 (t-test)를 진행하기 위해, 선행조건인 정규성 검정 진행.
- Python scipy library의 Anderson 모듈 이용. / 1_2.ipynb 64th cell 에서 진행
- 모든 교통안전물에 대해 test-statistic> critical value 이므로 정규분포 따르지 않음
- (아래수치에서 확인가능하며 왼쪽의 교통안전물 리스트 순)
- 따라서 t-test 검정이 불가하다.
- AndersonResult(statistic=166.73553957970853, critical_values=array([0.575, 0.655, 0.785, 0.916, 1.09]), significance_level=array([15. , 10. , 5. , 2.5, 1.]))
- AndersonResult(statistic=159.5647959916996, critical_values=array([0.575, 0.655, 0.785, 0.916, 1.09]), significance_level=array([15. , 10. , 5. , 2.5, 1.]))
- AndersonResult(statistic=174.61275943418923, critical_values=array([0.575, 0.654, 0.785, 0.916, 1.089]), significance_level=array([15. , 10. , 5. , 2.5, 1.]))
- AndersonResult(statistic=400.19992726004693, critical_values=array([0.575, 0.655, 0.786, 0.917, 1.091]), significance_level=array([15. , 10. , 5. , 2.5, 1.]))
- AndersonResult(statistic=175.36355077915596, critical_values=array([0.575, 0.655, 0.785, 0.916, 1.09]), significance_level=array([15. , 10. , 5. , 2.5, 1.]))
- AndersonResult(statistic=6.837172775656171, critical_values=array([0.539, 0.614, 0.736, 0.859, 1.022]), significance_level=array([15. , 10. , 5. , 2.5, 1.]))
- AndersonResult(statistic=493.4652430526003, critical_values=array([0.575, 0.655, 0.786, 0.917, 1.091]), significance_level=array([15. , 10. , 5. , 2.5, 1.]))
- AndersonResult(statistic=2.483355590293044, critical_values=array([0.549, 0.625, 0.75 , 0.875, 1.041]), significance_level=array([15. , 10. , 5. , 2.5, 1.]))
- AndersonResult(statistic=0.4235875237974227, critical_values=array([1.317, 1.499, 1.799, 2.098, 2.496]), significance_level=array([15. , 10. , 5. , 2.5, 1.]))

교통안전물 유무 구역의 사고횟수 밀도 함수



- 각 교통안전물이 위치한 구역과 그렇지 않은 구역의 사고횟수 밀도 함수 그래프
- x축: 구역당 발생한 사고횟수, y축: 발생한 구역 비율(수)
- 1_2.ipynb 65th cell 에서 진행 가능

모델 학습 loss 및 train data 성능

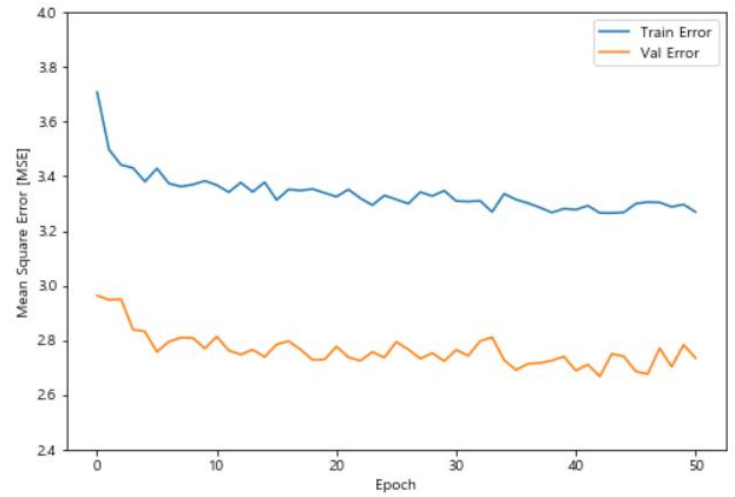


Train data 에 대한 예측 결과 값

실제 사고 지역에 대해 무사고 라고 예측할 확률: 12.98%
실제 무사고 지역에 대해 무사고 라고 예측할 확률: 90.82%
실제 (1회 이상) 사고 지역에 대해 사고 지역이라고 예측할 확률: 87.02%

실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 95.36%
실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 97.26%
실제 5회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 98.19%
실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.79%

모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 76.62%
모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 68.37%
모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 100.00%



Test data 에 대한 예측 결과 값

실제 사고 지역에 대해 무사고 라고 예측할 확률: 11.95%
실제 무사고 지역에 대해 무사고 라고 예측할 확률(specificity): 91.30%
실제 (1회 이상) 사고 지역에 대해 사고 지역이라고 예측할 확률(recall): 88.05%

실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 95.48%
실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 96.15%
실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%

모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 76.37%
모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 71.18%
모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 73.85%

선형 모델에 대한 성능 수치

Linear regression
- Python sklearn library의 linearRegression 모델 사용

Train data에 대한 모델 성능 평가

- 실제 사고 지역에 대해 무사고 라고 예측할 확률: 34.78%
- 실제 무사고 지역에 대해 무사고 라고 예측할 확률: 95.97%
- 모델이 사고 지역이라고 예측한 지역이 실제 사고 지역일 확률(precision):66.22%
- 실제 1회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 65.22%
- 실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 98.33%
- 실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.24%
- 실제 5회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.68%
- 실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%

- 모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 66.22%
- 모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 59.93%
- 모델이 3회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 55.74%
- 모델이 5회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 55.38%
- 모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 55.56%

test data에 대한 모델 성능 평가

- 실제 사고 지역에 대해 무사고 라고 예측할 확률: 36.20%
- 실제 무사고 지역에 대해 무사고 라고 예측할 확률(specificity): 96.28%
- 실제 (1회 이상) 사고 지역에 대해 사고 지역이라고 예측할 확률(recall): 63.80%
- 실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 97.74%
- 실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.15%
- 실제 5회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.15%
- 실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%

- 모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 67.56%
- 모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 55.15%
- 모델이 3회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 50.42%
- 모델이 5회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 47.17%
- 모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 47.37%

Ridge 선형회귀 모델에 대한 성능 수치

Ridge regression
- Python sklearn library의 Ridge 모델 사용

Train data에 대한 모델 성능 평가

- 실제 사고 지역에 대해 무사고 라고 예측할 확률: 34.78%
- 실제 무사고 지역에 대해 무사고 라고 예측할 확률: 95.97%
- 모델이 사고 지역이라고 예측한 지역이 실제 사고 지역일 확률(precision, 정밀도):66.22%
- 실제 1회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 65.22%
- 실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 98.33%
- 실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.24%
- 실제 5회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.68%
- 실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%

- 모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 66.22%
- 모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 59.93%
- 모델이 3회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 55.74%
- 모델이 5회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 55.38%
- 모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 55.56%

test data에 대한 모델 성능 평가

- 실제 사고 지역에 대해 무사고 라고 예측할 확률: 36.20%
- 실제 무사고 지역에 대해 무사고 라고 예측할 확률(specificity): 96.28%
- 실제 (1회 이상) 사고 지역에 대해 사고 지역이라고 예측할 확률(recall, 재현율): 63.80%
- 실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 97.74%
- 실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.15%
- 실제 5회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.15%
- 실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%

- 모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 67.56%
- 모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 55.15%
- 모델이 3회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 50.42%
- 모델이 5회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 47.17%
- 모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 47.37%

Lasso 선형회귀 모델에 대한 성능 수치

Ridge regression
- Python sklearn library의 Lasso 모델 사용

Train data에 대한 모델 성능 평가

- 실제 사고 지역에 대해 무사고 라고 예측할 확률: 33.84%
- 실제 무사고 지역에 대해 무사고 라고 예측할 확률: 95.79%
- 모델이 사고 지역이라고 예측한 지역이 실제 사고 지역일 확률(precision, 정밀도):65.58%
- 실제 1회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 66.16%
- 실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 98.45%
- 실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.15%
- 실제 5회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.92%
- 실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%

- 모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 65.58%
- 모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 59.82%
- 모델이 3회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 55.98%
- 모델이 5회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 55.70%
- 모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 56.49%

test data에 대한 모델 성능 평가

- 실제 사고 지역에 대해 무사고 라고 예측할 확률: 34.51%
- 실제 무사고 지역에 대해 무사고 라고 예측할 확률(specificity): 96.18%
- 실제 (1회 이상) 사고 지역에 대해 사고 지역이라고 예측할 확률(recall, 재현율): 65.49%
- 실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 97.18%
- 실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 98.72%
- 실제 5회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%
- 실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%

- 모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 67.53%
- 모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 54.97%
- 모델이 3회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 50.82%
- 모델이 5회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 47.12%
- 모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 44.44%

의사결정나무 모델에 대한 성능 수치

Decision Tree 회귀 모델
- Python sklearn library의 DecisionTreeRegressor 모델 사용
- Train data에 대해서만 제대로 학습하고, test data에 대해서는 성능이 매우 떨어지는 **과적합(overfitting)** 상태

Train data에 대한 모델 성능 평가

- 실제 사고 지역에 대해 무사고 라고 예측할 확률: 1.70%
- 실제 무사고 지역에 대해 무사고 라고 예측할 확률: 100.00%
- 모델이 사고 지역이라고 예측한 지역이 실제 사고 지역일 확률(precision, 정밀도):100.00%
- 실제 1회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 98.30%
- 실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%
- 실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%
- 실제 5회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%
- 실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%

- 모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 100.00%
- 모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 100.00%
- 모델이 3회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 100.00%
- 모델이 5회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 100.00%
- 모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 100.00%

test data에 대한 모델 성능 평가

- 실제 사고 지역에 대해 무사고 라고 예측할 확률: 40.74%
- 실제 무사고 지역에 대해 무사고 라고 예측할 확률(specificity): 95.59%
- 실제 (1회 이상) 사고 지역에 대해 사고 지역이라고 예측할 확률(recall, 재현율): 59.26%
- 실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 69.49%
- 실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 73.50%
- 실제 5회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 76.92%
- 실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 89.13%

- 모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 61.97%
- 모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 52.30%
- 모델이 3회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 46.15%
- 모델이 5회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 37.09%
- 모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 24.49%

랜덤포레스트 모델에 대한 성능 수치

Random forest 회귀 모델
- Python sklearn library의 RandomForestRegressor 모델 사용
- Train data에 대해서만 제대로 학습하고, test data에 대해서는 성능이 매우 떨어지는 **과적합(overfitting)** 상태

Train data에 대한 모델 성능 평가

- 실제 사고 지역에 대해 무사고 라고 예측할 확률: 21.70%
- 실제 무사고 지역에 대해 무사고 라고 예측할 확률: 99.09%
- 모델이 사고 지역이라고 예측한 지역이 실제 사고 지역일 확률(precision, 정밀도):91.28%
- 실제 1회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 78.30%
- 실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%
- 실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%
- 실제 5회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%
- 실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%

- 모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 91.28%
- 모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 90.23%
- 모델이 3회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 89.81%
- 모델이 5회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 89.56%
- 모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 90.98%

test data에 대한 모델 성능 평가

- 실제 사고 지역에 대해 무사고 라고 예측할 확률: 31.31%
- 실제 무사고 지역에 대해 무사고 라고 예측할 확률(specificity): 96.35%
- 실제 (1회 이상) 사고 지역에 대해 사고 지역이라고 예측할 확률(recall, 재현율): 68.69%
- 실제 2회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 99.15%
- 실제 3회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%
- 실제 5회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%
- 실제 10회 이상 사고 지역에 대해 사고 지역이라고 예측할 확률: 100.00%

- 모델이 1회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 69.51%
- 모델이 2회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 61.03%
- 모델이 3회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 56.05%
- 모델이 5회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 51.30%
- 모델이 10회 이상 사고날 것이라고 예측한 지역에 대해 실제로 맞을 확률: 38.46%

위험구역 가능성을 감소시키는 요인

- 평균적으로 위험구역의 가능성을 줄이는 요인
- Python lime library 이용하여 구현. 자세한 코드는 2_3.ipynb 에서 확인 가능 (df_ele)

표2. 위험구역 가능성을 **감소**시키는 요인

	중요도
건물연면적 <= 366366.72	-287.311448
차량등록대수 <= 0.00	-9.788768
0.00 < 차량등록대수 <= 1156.94	-3.654435
ic및jc <= 0.00	-3.479699
도로시설물 <= 0.00	-2.494872
교통량4 <= 419.57	-2.445083
교통량1 > 0.01	-1.575577
중앙분리대_개수 > 0.00	-1.019941
교차도로길이 > 0.00	-0.546938
정차금지지대_개수 <= 0.00	-0.457995
도로시설물 > 0.00	-0.409295
신호등_개수 > 2.05	-0.333336
1156.94 < 차량등록대수 <= 4590.25	-0.232639
0.00 < 교통안전표지_개수 <= 7.42	-0.179570
신호등_개수 <= 0.00	-0.136976
도로속도표시_개수 <= 0.00	-0.121999
교차로 > 0.00	-0.116888
횡단보도_개수 <= 0.00	-0.108446
7.42 < 교통안전표지_개수 <= 14.16	-0.072128

- Lime의 parameter: num_samples = 10000, num_features=10

교통안전물 사고 감소 효과

- 1회 이상 사고난 구역의, 각 교통안전물 추가시의 평균 예상되는 사고 감소 횟수
  - 횡단보도와 도로속도표시를 제외한 다른 교통안전물에 대해서는 오히려 사고가 증가.
  - 단순히 모든 구역에 교통안전물을 설치한다고 하여 효과가 있는 것은 아님.
  - 큰 수 의 사고 증가 예측은, 정말 사고가 증가한다는 것을 의미하는 것이 아니라 필요하지 않은 지역에 무리하게 해당 교통안전물을 설치해 모델이 잘못 인식했을 것으로 추측.

```
횡단보도_개수+1          1.600471
도로속도표시_개수+1       0.671508
교통안전표지_개수+1      -2.783434
교통cctv_개수+1      -7.700272
신호등_개수+1         -41.343612
중앙분리대_개수+1      -37.906310
안전지대_개수+1        -19.341094
정차금지지대_개수+1     -20.689628
```

위험지역 군집화 결과

- K-means clustering 결과, 각 군집의 중심점을 나타낸 자료.

1. 독립변수

	교차도로길이	횡단보도_개수	도로속도표시_개수	교통량4	ic및jc	교통안전표지_개수	차량등록대수	교통cctv_개수	교통량1	도로시설물	신호등_개수	건물연면적	총인구수	중앙분리대_개수	안전지대_개수	교차로	정차금지지대_개수	속성변화점
0	0.002	4.415	2.330	52.565	0.000	10.233	40.631	0.199	0.120	0.051	5.250	905.399	50.472	0.006	1.426	0.824	0.097	0.023
1	0.001	2.667	2.333	64.773	0.000	6.667	110.667	0.000	0.064	0.000	0.667	88318.970	143.667	0.000	0.667	0.000	0.333	0.000
2	0.002	3.333	2.000	33.976	0.000	8.167	57.500	0.333	0.061	0.167	3.000	33024.093	129.000	0.000	0.000	0.500	0.000	0.000
3	0.002	3.733	1.467	55.535	0.000	8.733	109.733	0.133	0.116	-0.000	3.267	10856.216	211.133	0.000	0.800	0.800	0.133	0.000

2. 사고 유형

	차대차 - 기타	차대차 - 추돌	차량단독	차대차 - 충돌	차대사람
0	0.068	0.379	0.068	0.583	0.621
1	0.971	1.088	0.265	3.118	0.294
2	1.381	0.254	0.079	0.175	0.159

3. 피해자 연령대

	0-10대	20대	3-50대	60대	고령
0	0.026	0.197	1.000	0.171	0.132
1	0.151	0.396	2.358	0.264	0.264
2	0.250	1.333	5.333	1.417	0.250
3	0.271	0.424	0.017	0.424	0.186

3. “An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning (2009)[26]

- 토지이용과 교통계획에 따라 교통사고 발생 양상이 다름
- 도로특성, 토지이용, 인구특성, 통근 수단에 따라 사고 모형을 log linear regression model로 설계
- 교통량이 차량대보행자 교통사고의 주요 원인
- 주거, 상업, 근린주거지역에서 차량 대 보행자 사고가 많이 발생