체감안전도 예측

COMPAS 러닝 머신 러닝 팀

시민이 공감하는 치안 체감안전도 예측

Machine Learning

Contents 체감안전도 예측

1. 목차

2. 데이터 전처리

결측치처리
- -범주형 변수
- -다항분포

구간화

위치데이터 처리

외부데이터 활용

one hot encoding

최종 데이터셋

3. 데이터선택

차원의 저주

해결법

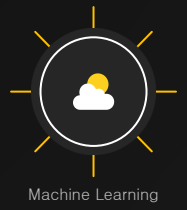
변수추출결과

4. 20년도 데이터 구성

5. 모델링

머신러닝
- -결과해석

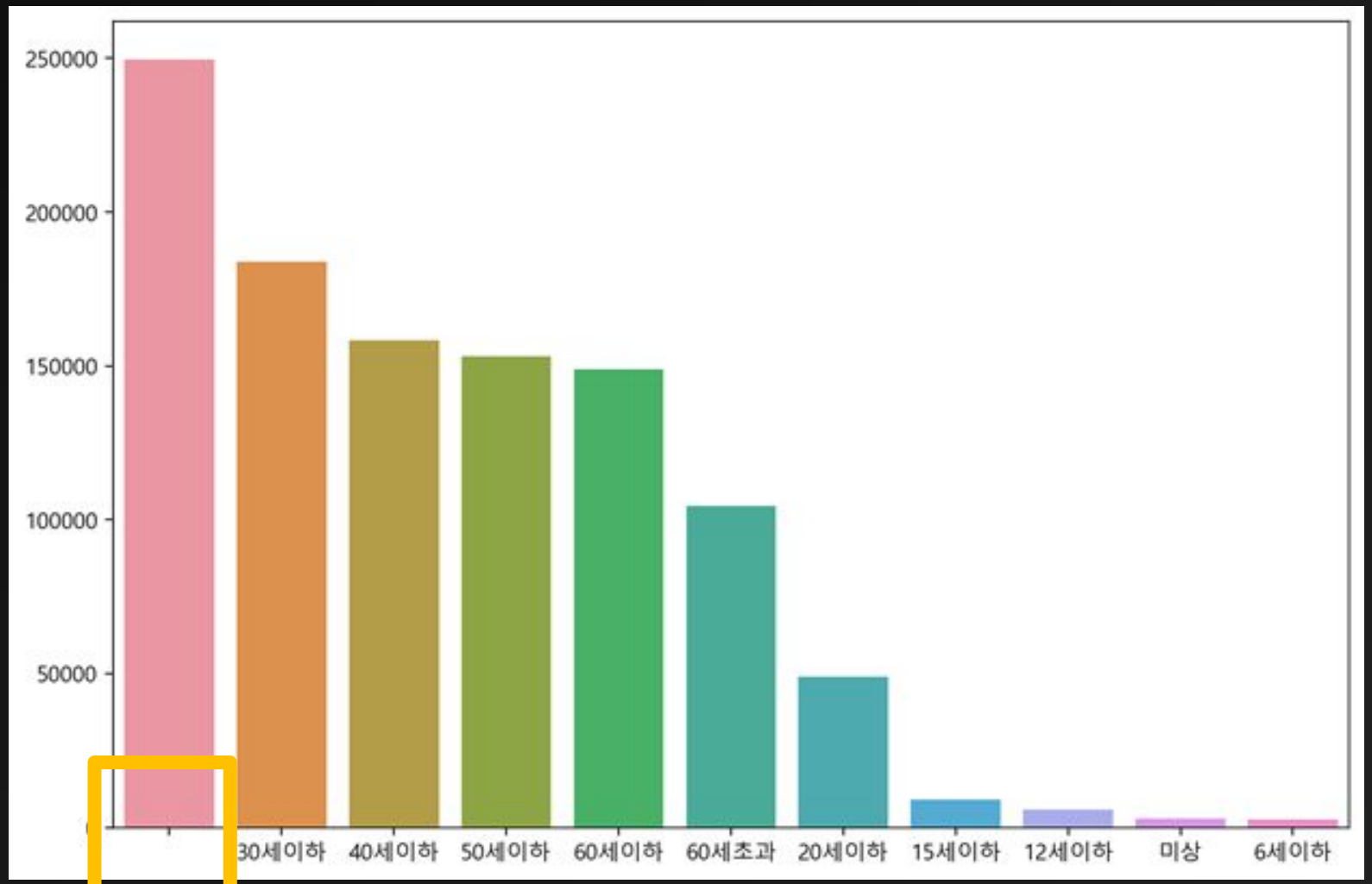
딥러닝
- -성능개선정도



데이터 전처리

결측치 처리

피해자 연령대 결측치 비율이 높아 제거하기 어려움 범주형 변수이므로 '미상'으로 표시



Data Preprocessing

체감안전도 예측-러닝머신러닝팀

Machine Learning

결측치 처리

피해자 연령대 결측치 비율이 높아 제거하기 어려움 범주형 변수이므로 '미상'으로 표시

	jur_stn	crm	vic_age
0	서울수서경찰서	위조외국통화행사	
1	서울영등포경찰서	도로교통법위반	
2	서울양천경찰서	209015100	60세초과
3	서울서초경찰서	폭행	40세이하
4	서울동대문경찰서	사기	30세이하
...	...	...	...
1068235	경남진해경찰서	사기	60세이하
1068236	경남진해경찰서	폭행	
1068237	경남마산동부경찰서	재물손괴	60세이하
1068238	경남마산중부경찰서	사기	50세이하
1068239	경남마산동부경찰서	강제추행	30세이하

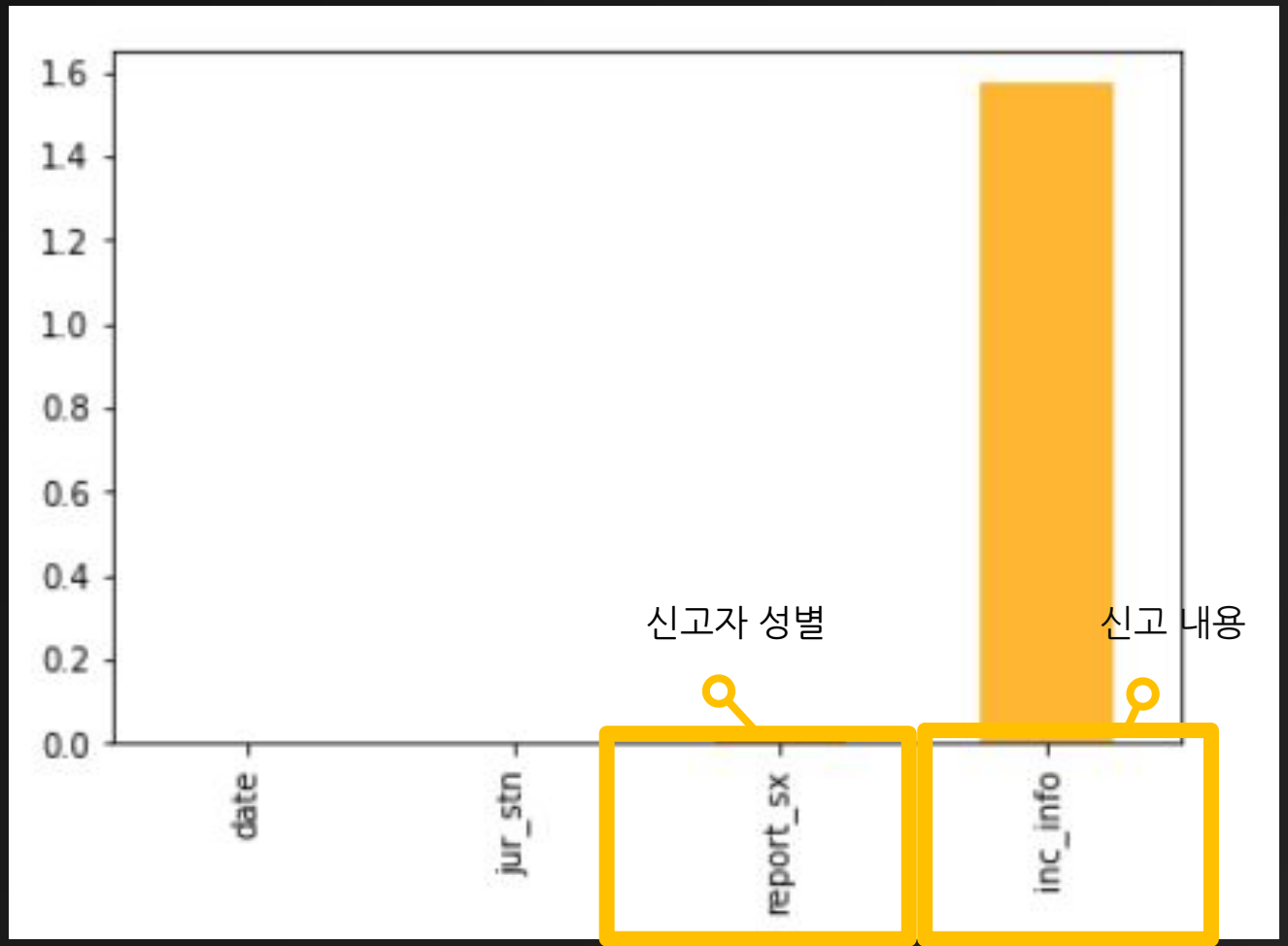
	jur_stn	crm	vic_age
0	서울수서경찰서	위조외국통화행사	미상
1	서울영등포경찰서	도로교통법위반	미상
2	서울양천경찰서	209015100	60세초과
3	서울서초경찰서	폭행	40세이하
4	서울동대문경찰서	사기	30세이하
...	...	...	...
1068235	경남진해경찰서	사기	60세이하
1068236	경남진해경찰서	폭행	미상
1068237	경남마산동부경찰서	재물손괴	60세이하
1068238	경남마산중부경찰서	사기	50세이하
1068239	경남마산동부경찰서	강제추행	30세이하

결측치 처리

112신고 파일

신고자의 성별과 신고 내용에 결측치

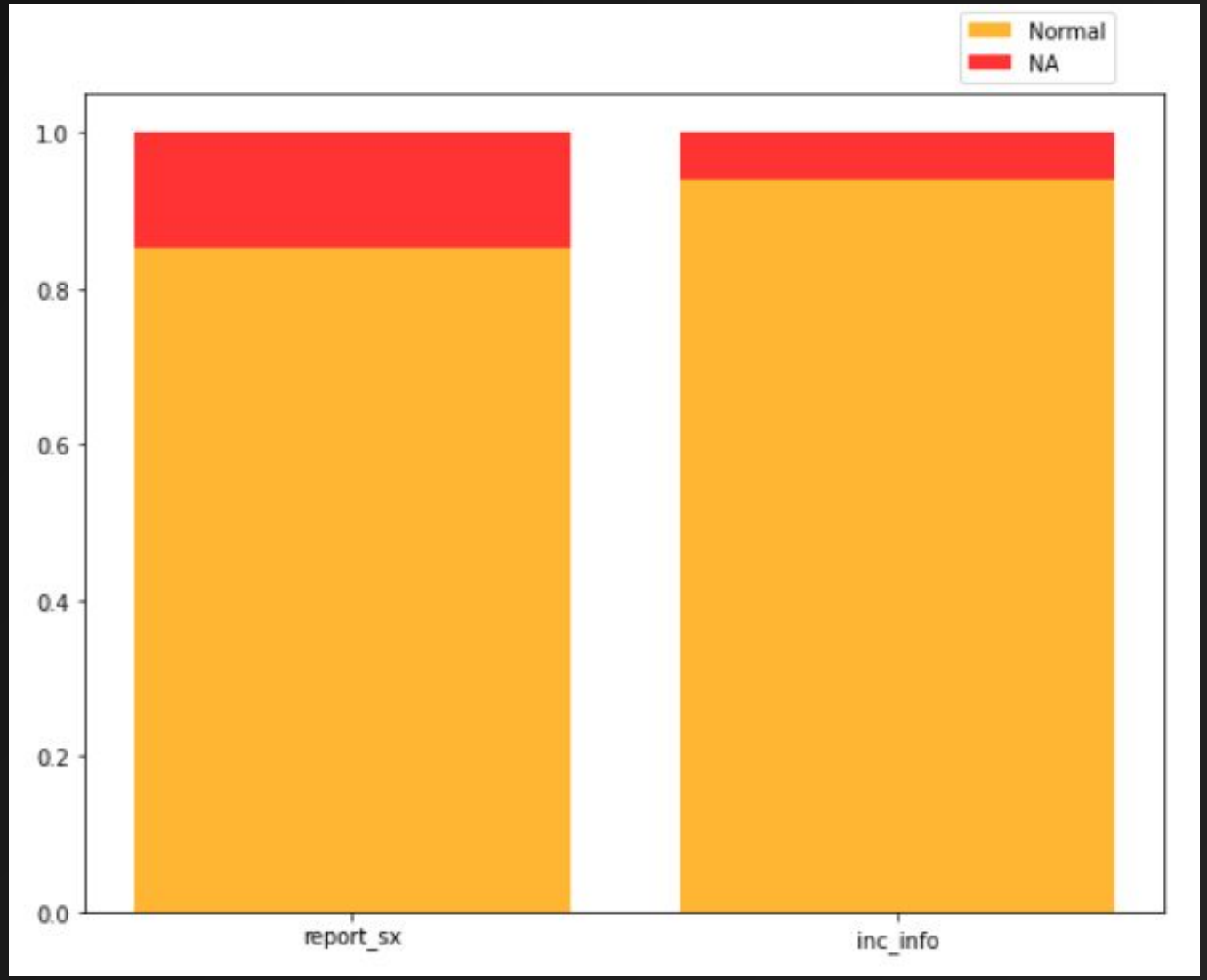
y 값은 전체 데이터에서 결측치의 백분위
신고자 성별은 0.2% 미만, 신고 내용은 1.6%
로의 매우 적은 결측치이므로 삭제 처리



결측치 처리

112신고 파일

미상, 내용확인 불가결측치
report_sx와 inc_info에서 각각 미상, 내용확인불가 라는 결측치가 존재
변수의 결측치 비율은 좌측 그래프



Data Preprocessing

체감안전도 예측-러닝머신러닝팀

Machine Learning

결측치 처리

112신고파일 2개의 유의미한 변수로 이루어져 있어, 두 변수 모두 결측치인 경우는 삭제

남성과 여성의 신고 내용은 다르게 나타날 것으로 예측되어, 남성과 여성의 내용 미상을 각기 다르게 처리

신고 성별 신고 내용

	date	jur_stn	report_sx	inc_info
51	20180603.0	서울용산	불상	내용확인불가
196	20180603.0	서울서대문	불상	내용확인불가
399	20180602.0	서울광진	불상	내용확인불가
539	20180603.0	서울성북	불상	내용확인불가
676	20180603.0	서울관악	불상	내용확인불가
...	...	...	...	...
9228978	20210531.0	진해	불상	내용확인불가
9228983	20210529.0	마산중부	불상	내용확인불가
9228988	20210531.0	진해	불상	내용확인불가
9229069	20210530.0	진해	불상	내용확인불가
9229073	20210530.0	진해	불상	내용확인불가

	date	jur_stn	report_sx	inc_info
122	2018	서울송파	남성	내용확인불가
1493	2018	서울송파	남성	내용확인불가
2962	2018	서울송파	남성	내용확인불가
3794	2018	서울송파	남성	내용확인불가
7427	2018	서울송파	여성	내용확인불가
...	...	...	...	...
1513913	2018	서울송파	여성	내용확인불가
1515220	2018	서울송파	여성	내용확인불가
1515980	2018	서울송파	여성	내용확인불가
1517859	2018	서울송파	남성	내용확인불가
1519293	2018	서울송파	여성	내용확인불가

결측치 처리

112신고 파일

대체값 추론

내용확인불가인 모집단에서 30번의 랜덤추출 후 평균값 계산. 이 값으로 결측치를 대체했고 모든 지역의 성별에 따른 112 신고의 내용 확인 불가 데이터를 같은 방법으로 처리.

구체적인 값의 내용은 그림과 같음

```
[[358 177 162 ...   1   0   0]
 [351 207 193 ...   0   0   0]
 [310 191 182 ...   0   0   0]
 ...
 [301 210 198 ...   0   0   0]
 [300 194 176 ...   0   0   0]
 [366 183 171 ...   0   0   0]]
[331.8666666666667, 198.26666666666668, 171.76666666666668, 126.46666666666667, 124.1]
```

```
inc_info
기타형사범    331.866667
보호조치      198.266667
상담문의      171.766667
시비        126.466667
교통사고      124.100000
소음         75.933333
교통불편       73.133333
폭력         71.766667
위험방지       58.900000
행패소란       53.166667
```

Data Preprocessing

체감안전도 예측-러닝머신러닝팀

Machine Learning

결측치 처리

112신고 파일

신고 내용을 전체 신고 수로 나누어 확률화

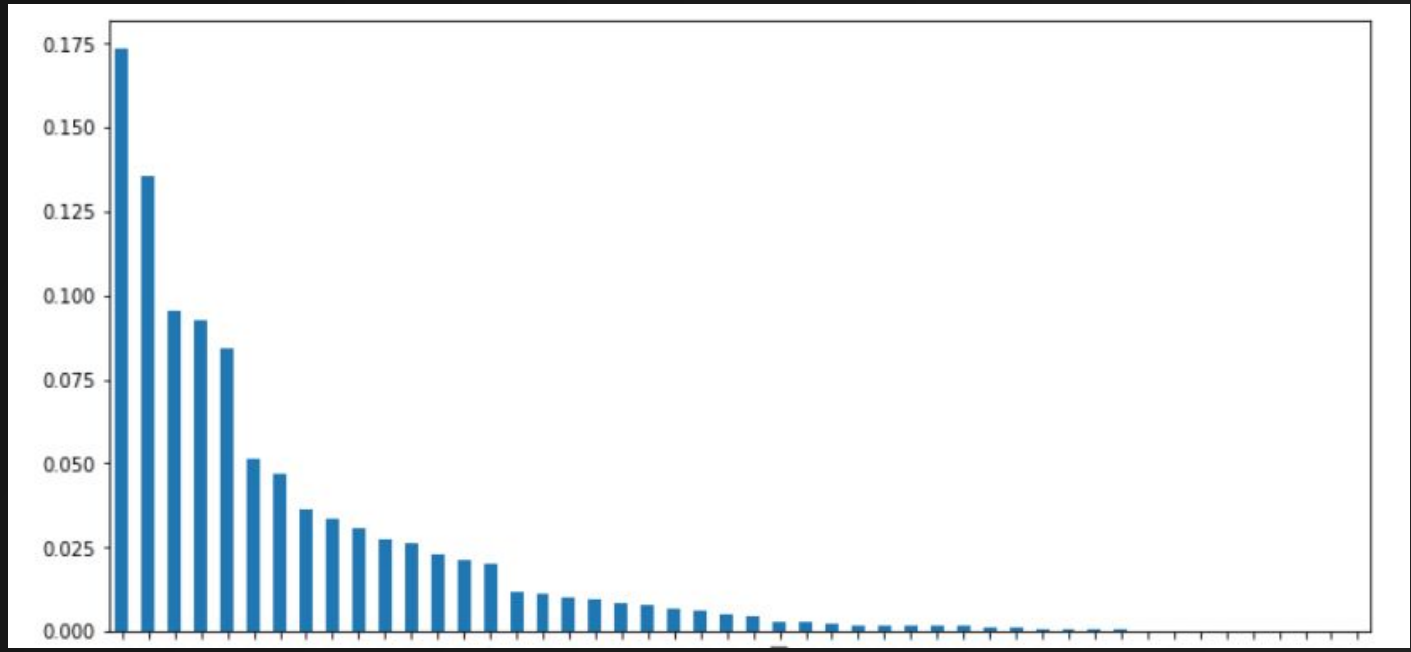
송파구 남성의 112 신고 전화를 확률변수라 가정

이 확률 변수는 결과가 48가지, 확률이 표와

그림과 같은 다항분포를 따름

처리하고자 하는 결측치, 신고내용 미상을 이

모집단에서 추출된 표본이라고 가정



inc_info	
기타형사범	0.191285
보호조치	0.115664
상담문의	0.101007
시비	0.074273
교통사고	0.071527
소음	0.042573
교통불편	0.042346
폭력	0.041606
위험방지	0.033824
행패소란	0.029814

Data Preprocessing

체감안전도 예측-러닝머신러닝팀

Machine Learning

결측치 처리

112신고 파일

결측치 처리가 끝난 데이터를5가지로 분류함

절도 폭력 : 폭력, 절도, 가정폭력, 데이트폭력, 성폭력, 사기, 학교폭력
강도 살인 : 변사자, 협박, 강도
교통사고 : 교통사고, 교통위반, 인피도주
질서 준수 : 시비, 소음, 교통불편, 교통위반, 행패소란, 분실습득
전반적 : 그 외
2018년은 하반기만 있으므로 **2배**로 늘림
2017년은 18년과 19년의 **평균치**로 대체

date	jur_stn	q1 신고
2018	서울수서경찰서	3528
2018	서울강남경찰서	6013
2018	서울종로경찰서	1337
2018	서울혜화경찰서	2206
2018	서울남대문경찰서	1977
2018	서울중부경찰서	2908
2018	서울방배경찰서	1299
2018	서울서초경찰서	4857
2018	서울서부경찰서	2335
2018	서울은평경찰서	3933

date	jur_stn	q1 신고
2018	서울수서경찰서	7056
2018	서울강남경찰서	12026
2018	서울종로경찰서	2674
2018	서울혜화경찰서	4412
2018	서울남대문경찰서	3954
2018	서울중부경찰서	5816
2018	서울방배경찰서	2598
2018	서울서초경찰서	9714
2018	서울서부경찰서	4670
2018	서울은평경찰서	7866

date	jur_stn	q1 신고
2018	서울수서경찰서	7056
2018	서울강남경찰서	12026
2018	서울종로경찰서	2674
2018	서울혜화경찰서	4412
2018	서울남대문경찰서	3954
2018	서울중부경찰서	5816
2018	서울방배경찰서	2598
2018	서울서초경찰서	9714
2018	서울서부경찰서	4670
2018	서울은평경찰서	7866

date	jur_stn	q1 신고
2019	서울수서경찰서	7312
2019	서울강남경찰서	10084
2019	서울종로경찰서	4663
2019	서울혜화경찰서	3499
2019	서울남대문경찰서	3451
2019	서울중부경찰서	4714
2019	서울방배경찰서	2077
2019	서울서초경찰서	9423
2019	서울서부경찰서	4162
2019	서울은평경찰서	6736

date	jur_stn	q1 신고
2017	서울수서경찰서	7184
2017	서울강남경찰서	11055
2017	서울종로경찰서	3668
2017	서울혜화경찰서	3956
2017	서울남대문경찰서	3702
2017	서울중부경찰서	5265
2017	서울방배경찰서	2338
2017	서울서초경찰서	9568
2017	서울서부경찰서	4416
2017	서울은평경찰서	7301

외부데이터 인용

1인가구수 데이터

연도 부재 주어진 데이터가 자치구 기준으로 되어있는 반면, 몇몇 관할서는 읍면동 기준으로 관할 구역이 나누어져 있으므로, KOSIS에서 읍면동 기준 1인 가구 데이터를 이용해 전처리 진행함.
15년과 20년 데이터만 존재하므로, 인구 증가가 일정한 선형으로 이루어진다고 가정하고20년과 15년 인구의 차를 5로 나누어 17,18,19년도 인구를 계산

행정구역별(읍면동)	2015	2015	2020	2020
행정구역별(읍면동)	일반가구_계	1인	일반가구_계	1인
반포본동	3717	265	3601	354
반포2동	5592	499	4591	314
방배본동	7141	1475	7308	1683
방배1동	6480	1914	6973	2318
방배2동	10111	2594	7708	2182
방배3동	7866	1359	7407	1472
방배4동	9295	2407	8917	2585

	연도	총 1인가구 수
서울방배경찰서	2017	10671.0
서울방배경찰서	2018	10750.0
서울방배경찰서	2019	10829.0

구간화

각 데이터별 가독성 향상을 위한 시간 및 나이 데이터 구간화

미상, 00:00~02:59, 03:00~05:59, 06:00~08:59, 09:00~11:59, 12:00-14:59, 15:00~17:59, 18:00~20:59, 21:00~23:59로 나뉘어져 있지만

미상, 새벽, 오전, 오후, 저녁 5개 구간으로 구간화

	jur_stn	crm	crm_wthr	crm_clue	vic_sx	vic_age	crm_loc	crm_tm
0	서울수서경찰서	위조외국통화행사	미상	진정	불상		은행	09:00-11:59
1	서울영등포경찰서	도로교통법위반	맑음	타인신고	불상		노상	21:00-23:59
2	서울양천경찰서	209015100	미상	피해자신고	남자	60세초과	노상	미상
3	서울서초경찰서	폭행	미상	피해자신고	여자	40세이하	기타	21:00-23:59
4	서울동대문경찰서	사기	미상	진정	여자	30세이하	기타	미상

crm_tm
오전
저녁
미상
저녁
미상
...
미상
새벽

구간화

각 데이터별 가독성 향상을 위한 시간 및 나이 데이터 구간화

시도, 구별로 연령별 인원수로 구성된 데이터를 통합하여 나이대별(15세 이하, 20 ~ 34세, 35~ 59세, 60세 이상) 인원수로 구간화

	date	sido	sgg_nm	age	num
0	201706	서울특별시	종로구	4	4
1	201706	서울특별시	종로구	45	26
2	201706	서울특별시	종로구	46	55
3	201706	서울특별시	종로구	47	37
4	201706	서울특별시	종로구	48	51
...	...	...	...	...	...
145466	202105	경상남도	진주시	39	94
145467	202105	경상남도	진주시	40	122
145468	202105	경상남도	진주시	41	135
145469	202105	경상남도	진주시	42	114
145470	202105	경상남도	진주시	43	109

145471 rows × 5 columns

	police	year	19세 이하	20 ~ 34세	35 ~ 59세	60세 이상	총합
0	경기수원남부경찰서	2017	39101	9131	40639	50273	139144
1	경기수원서부경찰서	2017	39101	9131	40639	50273	139144
2	경기수원중부경찰서	2017	39101	9131	40639	50273	139144
3	경남마산동부경찰서	2017	54889	11055	56994	66509	189447
4	경남마산중부경찰서	2017	54889	11055	56994	66509	189447
...	...	...	...	...	...	...	...
159	서울종암경찰서	2020	30625	13009	48861	79450	171945
160	서울중랑경찰서	2020	46362	19807	80331	116848	263348
161	서울중부경찰서	2020	6478	2912	16092	33159	58641
162	서울혜화경찰서	2020	5516	2985	16072	28596	53169
163	충남세종경찰서	2020	19178	5176	20354	27770	72478

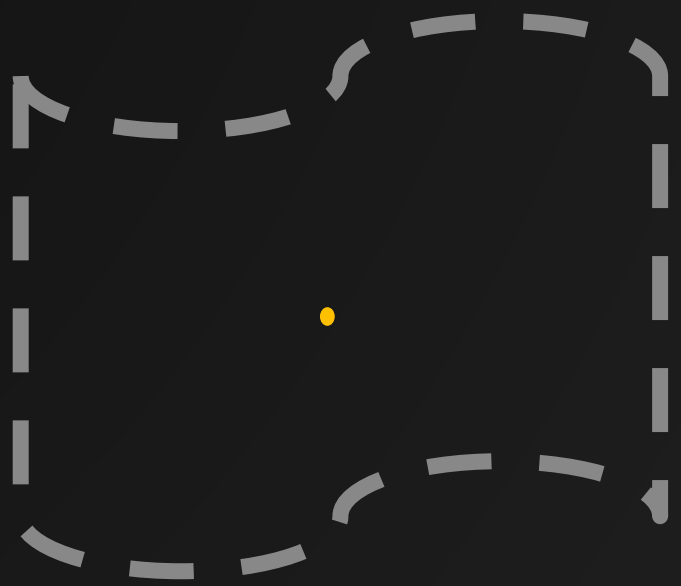
164 rows × 7 columns

위치데이터처리

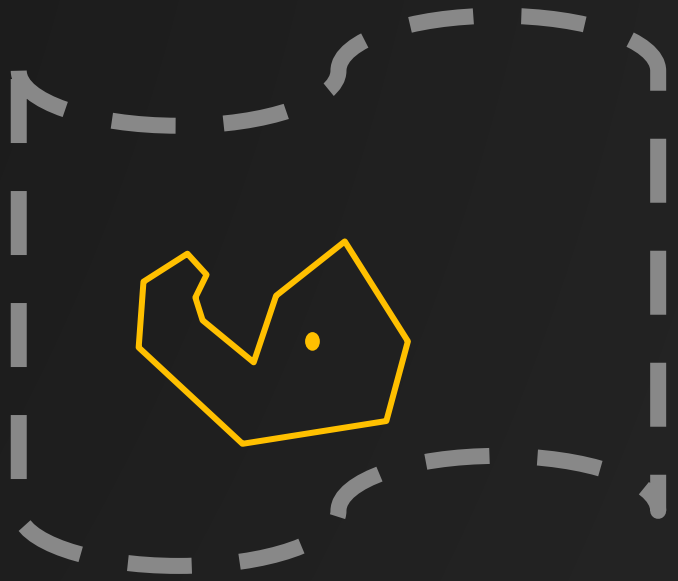
관할서 파생변수 생성

데이터 전처리 위도경도 정보를 통해 관할서 지정

Geopandas로 관할서 경계를 나타내는 json파일의 polygon을 표현



위도경도 데이터를 shapely.geometry의 point()함수를 통해 지표위의 한 점 표시



지도위에 표시한 점이 관할서 경계안에 있다면 경계에 안에 있는 위도, 경도 데이터에 해당 관할서명 추가

Data Preprocessing

체감안전도 예측-러닝머신러닝팀

Machine Learning

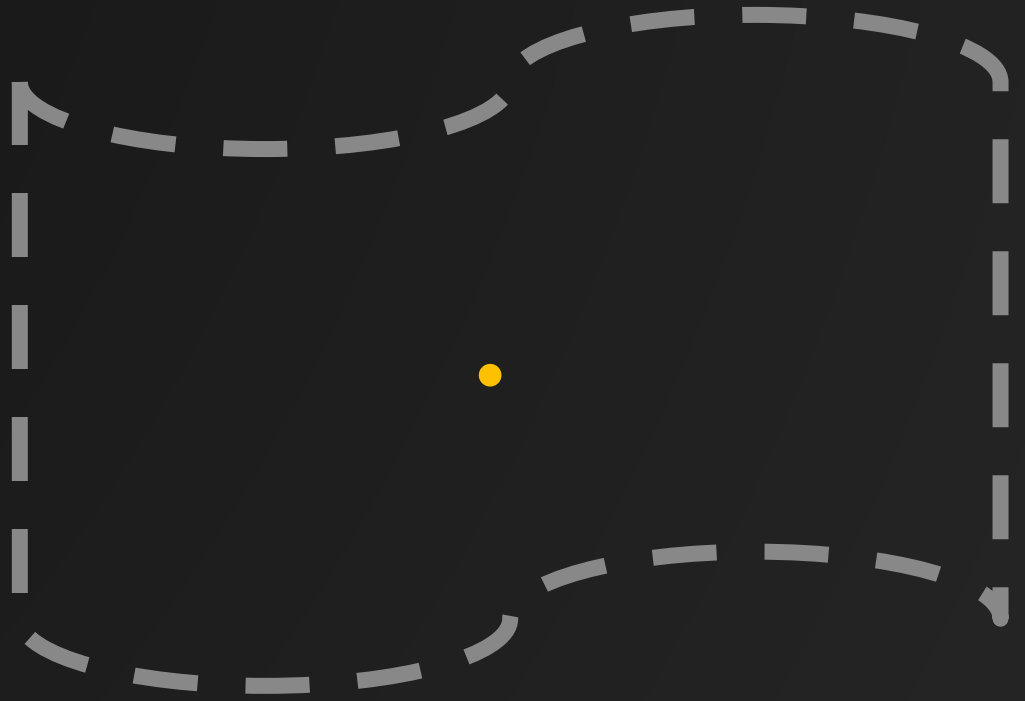
위치데이터처리

관할서 파생변수 생성

데이터 전처리 위도경도 정보를 통해 관할서 지정

	NAME	PNAME	geometry
0	세종경찰서	충남청	MULTIPOLYGON (((127.17202 36.73106, 127.17202 ...
1	진주경찰서	경남청	MULTIPOLYGON (((128.26697 35.12927, 128.26697 ...
2	창원서부경찰서	경남청	MULTIPOLYGON (((128.63363 35.22152, 128.63357 ...
3	창원중부경찰서	경남청	MULTIPOLYGON (((128.60966 35.15093, 128.60956 ...
4	마산동부경찰서	경남청	MULTIPOLYGON (((128.62696 35.21714, 128.62695 ...

위도경도 데이터를 shapely.geometry의 point()
함수를 통해 지표위의 한 점으로 표현



위치데이터처리

관할서 파생변수 생성

데이터 전처리 위도경도 정보를 통해 관할서 지정

	NAME	PNAME	geometry
0	세종경찰서	충남청	MULTIPOLYGON (((127.17202 36.73106, 127.17202 ...
1	진주경찰서	경남청	MULTIPOLYGON (((128.26697 35.12927, 128.26697 ...
2	창원서부경찰서	경남청	MULTIPOLYGON (((128.63363 35.22152, 128.63357 ...
3	창원중부경찰서	경남청	MULTIPOLYGON (((128.60966 35.15093, 128.60956 ...
4	마산동부경찰서	경남청	MULTIPOLYGON (((128.62696 35.21714, 128.62695 ...

Geopandas로 관할서 경계를 나타내는 json파일의 polygon을 표현

Data Preprocessing

체감안전도 예측-러닝머신러닝팀

Machine Learning

위치데이터처리

관할서 파생변수 생성

데이터 전처리 위도경도 정보를 통해 관할서 지정

	NAME	PNAME	geometry
0	세종경찰서	충남청	MULTIPOLYGON (((127.17202 36.73106, 127.17202 ...
1	진주경찰서	경남청	MULTIPOLYGON (((128.26697 35.12927, 128.26697 ...
2	창원서부경찰서	경남청	MULTIPOLYGON (((128.63363 35.22152, 128.63357 ...
3	창원중부경찰서	경남청	MULTIPOLYGON (((128.60966 35.15093, 128.60956 ...
4	마산동부경찰서	경남청	MULTIPOLYGON (((128.62696 35.21714, 128.62695 ...

	address	securitylight_cnt	lon	lat	관할서
0	세종특별자치시 금남면 감성리 64-2	1	127.287690	36.443467	세종경찰서
1	세종특별자치시 금남면 감성리 267	1	127.288812	36.444181	세종경찰서
2	세종특별자치시 금남면 감성리 40-1	1	127.289575	36.444711	세종경찰서
3	세종특별자치시 금남면 감성리 26	1	127.290071	36.444455	세종경찰서
4	세종특별자치시 금남면 감성리 267	1	127.290002	36.444188	세종경찰서

Data Preprocessing

체감안전도 예측-러닝머신러닝팀

Machine Learning

범주형 변수

One Hot encoding

범주형 변수를 One Hot encoding하여 모델링

	crm_wthr_눈	crm_wthr_만월	crm_wthr_맑음	crm_wthr_미상	crm_wthr_바람	crm_wthr_비	crm_wthr_안개	crm_wthr_암흑	crm_wthr_폭설	crm_wthr_폭풍우	...	vic_age_10대	vic_age_2,30대	v
0	0	0	0	1	0	0	0	0	0	0	...	0	1	
1	0	0	1	0	0	0	0	0	0	0	...	0	1	
2	0	0	1	0	0	0	0	0	0	0	...	0	1	
3	0	0	1	0	0	0	0	0	0	0	...	0	1	
4	0	0	1	0	0	0	0	0	0	0	...	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
360011	0	0	1	0	0	0	0	0	0	0	...	0	0	
360012	0	0	1	0	0	0	0	0	0	0	...	0	1	
360013	0	0	0	1	0	0	0	0	0	0	...	0	0	
360014	0	0	1	0	0	0	0	0	0	0	...	0	0	
360015	0	0	0	1	0	0	0	0	0	0	...	0	0	

360016 rows × 40 columns

Data Preprocessing

체감안전도 예측-러닝머신러닝팀

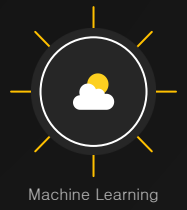
Machine Learning

최종데이터셋

관할서별, 연도별로 나눈

총 123개의 데이터 셋

	관할서	연도	crm_wthr_눈	crm_wthr_만월	crm_wthr_맑음	crm_wthr_미상	crm_wthr_바람	crm_wthr_비	crm_wthr_안개	crm_wthr_암흑	...	자살_사망률_10만명당	자살_연령표준화사망률_10만명당	총_인구수	기초수급_19세이하	기초수급_20_34세	기초수급_35_59세	기초수급_60세이상	기초수급_종합	외국인수	score_절욕
118	서울종암경찰서	2019	1.0	1.0	702.0	599.0	0.0	23.0	0.0	43.0	...	13.5	11.3	191004	31361	11882	43894	69254	156391	7111	75.80
119	서울동작경찰서	2019	3.0	0.0	2051.0	1395.0	0.0	49.0	0.0	225.0	...	17.6	15.4	386331	46485	15246	64943	92574	219248	9785	71.85
120	서울동부경찰서	2019	2.0	0.0	1475.0	295.0	0.0	67.0	0.0	170.0	...	8.1	8.8	103318	6833	2472	14862	28268	52435	11496	77.40
121	서울혜화경찰서	2019	0.0	0.0	999.0	265.0	1.0	32.0	0.0	30.0	...	12.0	11.4	75332	5756	2573	14071	25508	47908	10641	74.60
122	충남세종경찰서	2019	0.0	0.0	559.0	759.0	1.0	9.0	0.0	87.0	...	18.4	18.2	338136	16060	4101	15794	22678	58633	9814	80.05



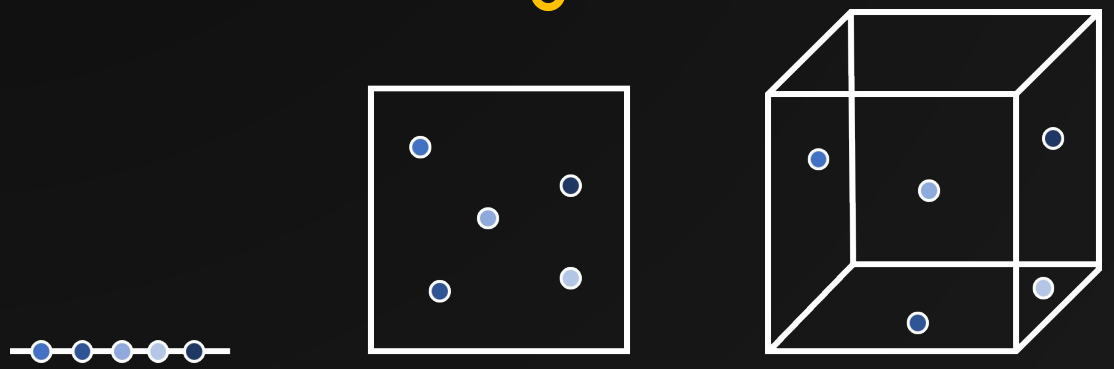
데이터 선택

Selection of Data 체감안전도 예측-러닝머신러닝팀

Machine Learning

차원의 저주

- 차원(변수의 수)이 증가함에 따라 성능이 저하되는 현상
- 특히 관측치의 수보다 변수의 수가 많아지면 발생



1차원 2차원 3차원

차원의 저주가 발생하면 성능이 저하되는 이유

- 차원이 증가하면서 데이터수가 급격하게 적어짐
- 모델학습과 추론의 계산복잡도가 높아짐

121	서울혜화경찰서	2019	1.0	0.0	660.0	1.0	0.0	27.0	0.0
122	충남세종경찰서	2019	2.0	0.0	913.0	201.0	0.0	73.0	3.0

123 rows × 131 columns

관측치의 수: 123개의 관할서
변수의 수: 131개
차원의 저주 발생

Selection of Data

체감안전도 예측-러닝머신러닝팀

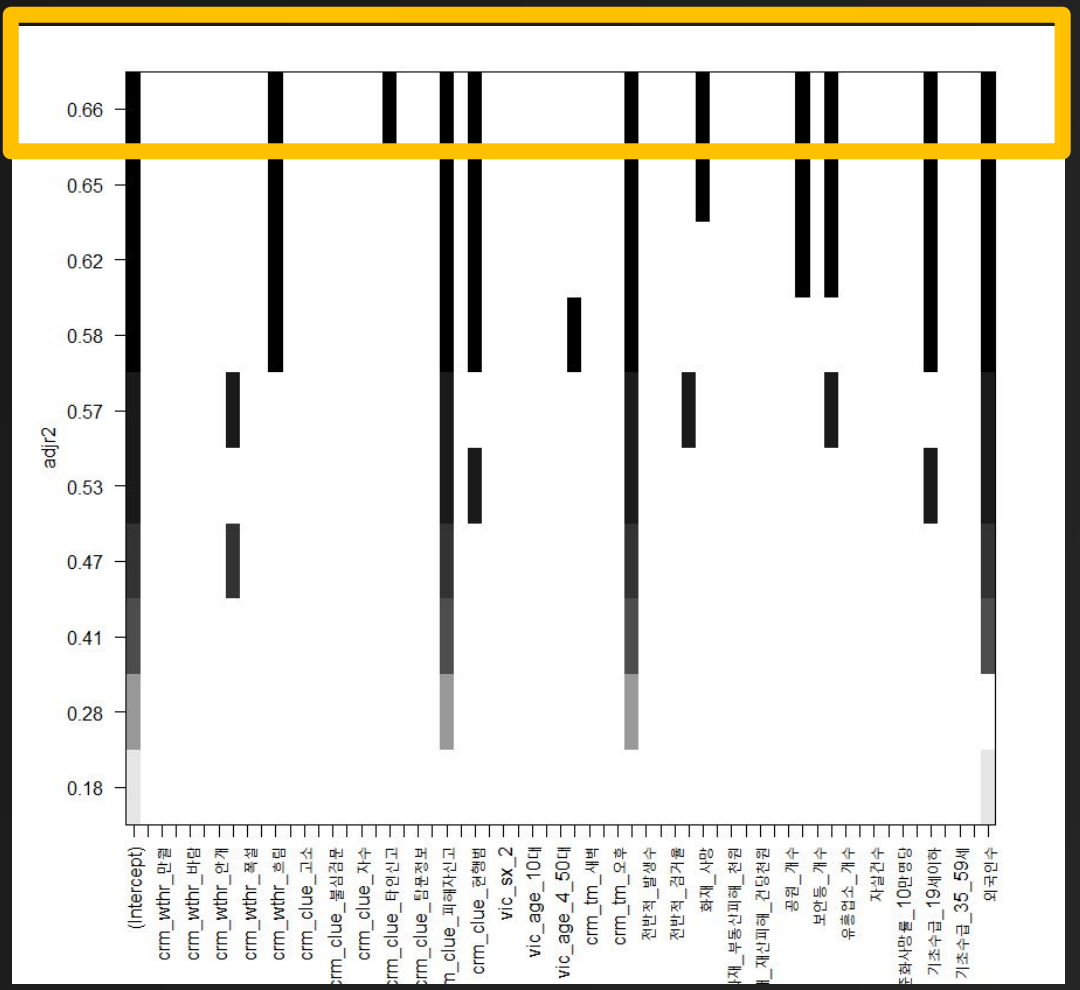
Machine Learning

차원의 저주

해결법

- 차원의 저주를 해결하는 방법: feature selection
- 전진선택, 후진소거법은 단계적으로 추가/삭제 하기에 최고의 성능을 낼 수 없다.
- 이를 보완한 것이 R의 All Subset Regression
- R-All Subset Regression
  n개의 설명변수를 추가하거나 뺀 2^n의 회귀모델을 만들고
  이들을 비교하여 가장 설명력이 높은 모델을 선택

절도폭력 안전도 데이터셋 에서는
가장 높은 설명력(0.66)을 나타내는 10개의 변수가 채택



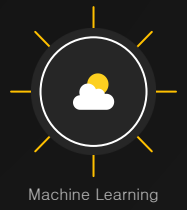
Selection of Data

체감안전도 예측-러닝머신러닝팀

Machine Learning

변수 추출 결과

절도폭력안전도	crm_wthr_바람,crm_clue_피해자신고,crm_clue_고소,crm_clue_현행범,vic_sx_2,vic_age_60세초과,cctv_개수,배치인원_수,비상벨_개수,일인가구수,기초수급_19세이하,외국인수
강도살인안전도	crm_clue_변사체,crm_clue_자수,crm_clue_진정,crm_clue_현행범,crm_tm_저녁,vic_sx_1,강도살인_검거수,화재_사망,화재_부상,화재_부동산피해_천원,공원_개수,자살_사망률_10만명당,자살_연령표준화사망률_10만명당,기초수급_35_59세,기초수급_60세이상,외국인수
교통사고안전도	crm_wthr_눈,crm_clue_피해자신고,vic_age_2_30대,일인가구수,총_인구수,기초수급_19세이하,기초수급_60세이상,외국인수,crm_clue_자수,crm_clue_타인신고,crm_clue_현행범,crm_clue_탐문정보,vic_sx_2,기초수급_19세이하,기초수급_20_34세,crm_tm_새벽,화재_사망,cctv_개수,배치인원_수,비상벨_개수,일인가구수
법질서 안전도	crm_clue_자수,crm_clue_타인신고,crm_clue_현행범,crm_clue_탐문정보,vic_sx_2,기초수급_19세이하,기초수급_20_34세,crm_tm_새벽,화재_사망,cctv_개수,배치인원_수,비상벨_개수,일인가구수
전반적안전도	crm_wthr_흐림,crm_clue_타인신고,crm_clue_피해자신고,crm_clue_현행범,crm_tm_저녁,화재_사망,배치인원_수,비상벨_개수,기초수급_19세이하,외국인수



20년도 데이터 구성

Dataset Deployment

체감안전도 예측-러닝머신러닝팀

Machine Learning

테스트셋 구성

- 범죄발생원표의 범죄발생단서에 대한 2020년도 데이터셋은 없기 때문에 논리적인 방법으로 예측해야 한다.
- 경찰청의 경찰범죄통계에서 범죄발생원표의 범죄발생단서에 해당하는 2020년도 데이터를 이용한다
- 경찰청에서 제공하는 데이터는 전국단위의 2020년도 데이터 이므로 관할서별 데이터 분류가 불가능하다
- 따라서 전국단위의 2020년도 데이터를 모집단으로 가정하고 19년도 전국 데이터와 20년도 전국 데이터의 증감율을 대입하여 각 관할서별 2020년도 데이터셋을 구축한다.

1) 범죄의 수사단서

자료갱신일: 2021-08-02 / 수록기간: 년 2011 ~ 2020 / 자료문의처 : 02-3150-1632

일괄설정 | 항목[1/1] | 죄종별[44/44] | 수사단서별[16/16] | 시점[1/10]

(단위 : 건) 주석정보 주소정보 행열전환

죄종별(1)	죄종별(2)	2020 계	현행범	신고			
		소계	소계	소계	피해자신고	고소	고발
계	소계	1,587,866	83,585	1,288,087	631,733	267,531	46,038
강력범죄	소계	24,332	2,750	20,774	9,490	8,161	91
	살인(기수)	308	60	160	21	19	2
	살인(미수등)	416	162	235	76	47	6
	강도	663	92	468	333	70	-
	강간	5,313	293	4,831	1,576	2,593	24
	유사강간	823	37	760	255	377	5
	강제추행	15,344	1,682	13,321	6,679	4,954	52
	3) 기타 강간·강제추행등	237	15	207	78	86	-
	방화	1,228	409	792	472	15	2
절도범죄	소계	179,517	4,268	159,232	150,505	3,619	110
폭력범죄	소계	265,768	35,405	226,775	175,506	31,268	355
	상해	32,904	4,262	28,101	17,609	7,314	46
	폭행	139,361	20,916	116,965	95,832	11,716	88
	체포·감금	1,209	192	994	502	356	9
	협박	21,214	3,662	17,183	10,673	5,138	38
	약취·유인	210	8	190	75	70	4
	폭력행위등	7,398	656	6,377	4,045	1,363	26

	A	B	C	E
1	변수	2019	2020	19년도 대비 20년도 증감율
2	crm_clue_고소	34,965	34,887	0.99777
3	crm_wthr_바람			
4	crm_clue_현행범	53,794	39,673	0.73750
5	crm_clue_피해자신고	340,412	326,011	0.95770
6	vic_sx_2	163,735	151,580	0.92576
7	vic_age_60세초과	61752	61970	1.00353

경찰통계 사이트 2020년도 범죄발생단서 데이터

19년도 대비 20년도 증감율 조사 데이터

Dataset Deployment

체감안전도 예측-러닝머신러닝팀

Machine Learning

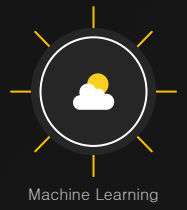
테스트셋 구성

	관할서	연도	crm_clue_고소	crm_wthr_바람	crm_clue_현행범	crm_clue_피해자신고	vic_sx_2	vic_age_60세초과
0	경기수원남부경찰서	2019	186.0	0.0	605.0	3923.0	1745.0	330.0
1	경기수원서부경찰서	2019	76.0	0.0	340.0	2565.0	1101.0	354.0
2	경기수원중부경찰	2019	126.0	0.0	494.0	2633.0	1347.0	493.0

E	
19년도 대비 20년도 증감율	
["crm_clue_고소"]	0.99777
["crm_clue_현행범"]	0.73750
["crm_clue_피해자신고"]	0.95770
["vic_sx_2"]	0.92576
["vic_age_60세초과"]	1.00353

[데이터]×[증감율] =

	관할서	연도	crm_clue_고소	crm_wthr_바람	crm_clue_현행범	crm_clue_피해자신고	vic_sx_2	vic_age_60세초과
0	경기수원남부경찰서	2020	186.0	0.0	446.0	3757.0	1615.0	331.0
1	경기수원서부경찰서	2020	76.0	0.0	251.0	2457.0	1019.0	355.0
2	경기수원중부경찰서	2020	126.0	0.0	364.0	2522.0	1247.0	495.0



모델링 [머신러닝]

모델링 방법 및 순서

Scaling

- non-scaling을 포함한 성능향상을 위해 4가지의 데이터 스케일링 작업
- 아래 모델리스트에 있는 알고리즘들을 적용
- 정확한 오차점수 도출을 위한 k-fold교차검증 적용
- 그 중 mae값이 가장 낮은 모델 선택

Non-scaling	Minmax-scaling	Standard-scaling	Robust-scaling
Linear regression	Linear regression	Linear regression	Linear regression
Ridge	Ridge	Ridge	Ridge
Lasso	Lasso	Lasso	Lasso
Elastic	Elastic	Elastic	Elastic
Xgboost	Xgboost	Xgboost	Xgboost
Lightbm	Lightbm	Lightbm	Lightbm
	Support vector regression	Support vector regression	Support vector regression

모델링 방법 및 순서

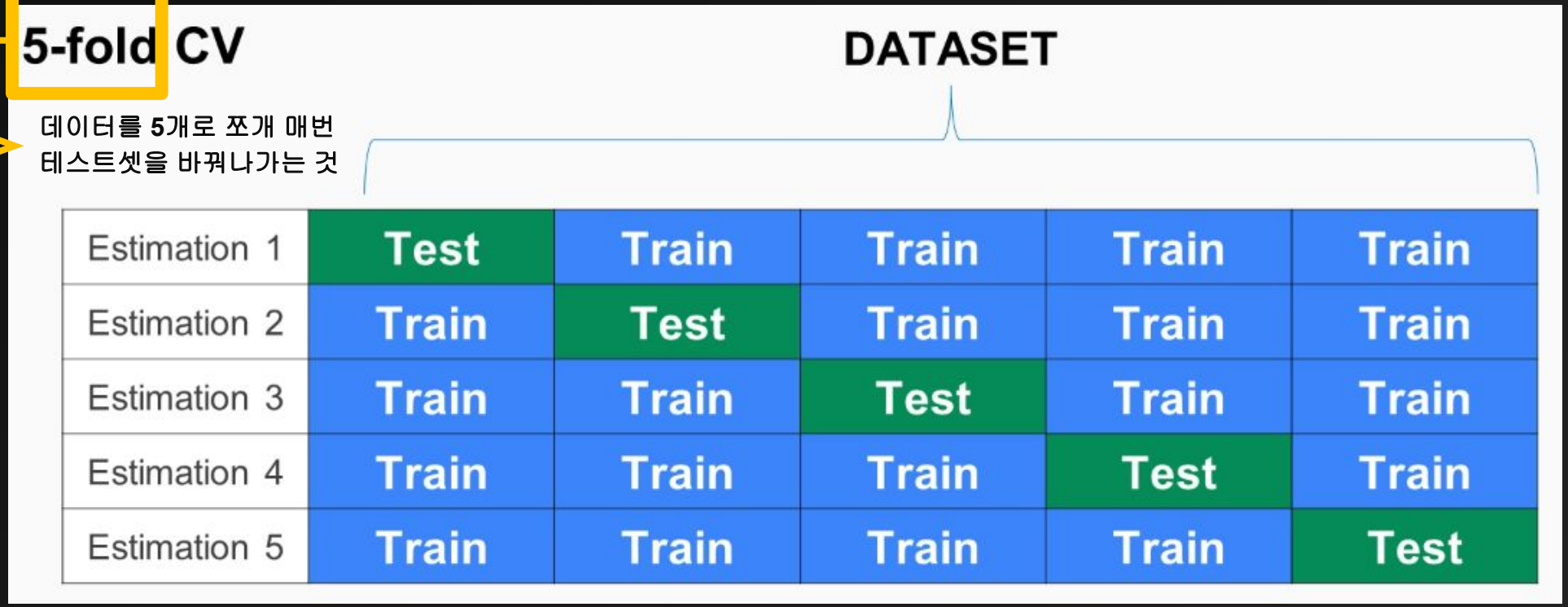
K-fold 교차검증

- K-겹 교차 검증은 모든 데이터가 최소 한 번은 테스트셋으로 쓰임.
- 총 데이터수가 적은 데이터셋에 대해 성능평가 정확도 향상 가능
- 정확한 오차점수 도출을 위해 k-fold교차검과적합을 방지하여 성능평가 정확도 향상 가능

∴ 데이터수가 123개인 현 상황에선 최적의 검증방법

5-fold CV

데이터를 5개로 쪼개 매번 테스트셋을 바꿔나가는 것



출처:
https://subscription.packtpub.com/book/big-data-and-business-intelligence/9781789617740/2/ch02lvl1sec14/k-fold-cross-validation

모델링 결과

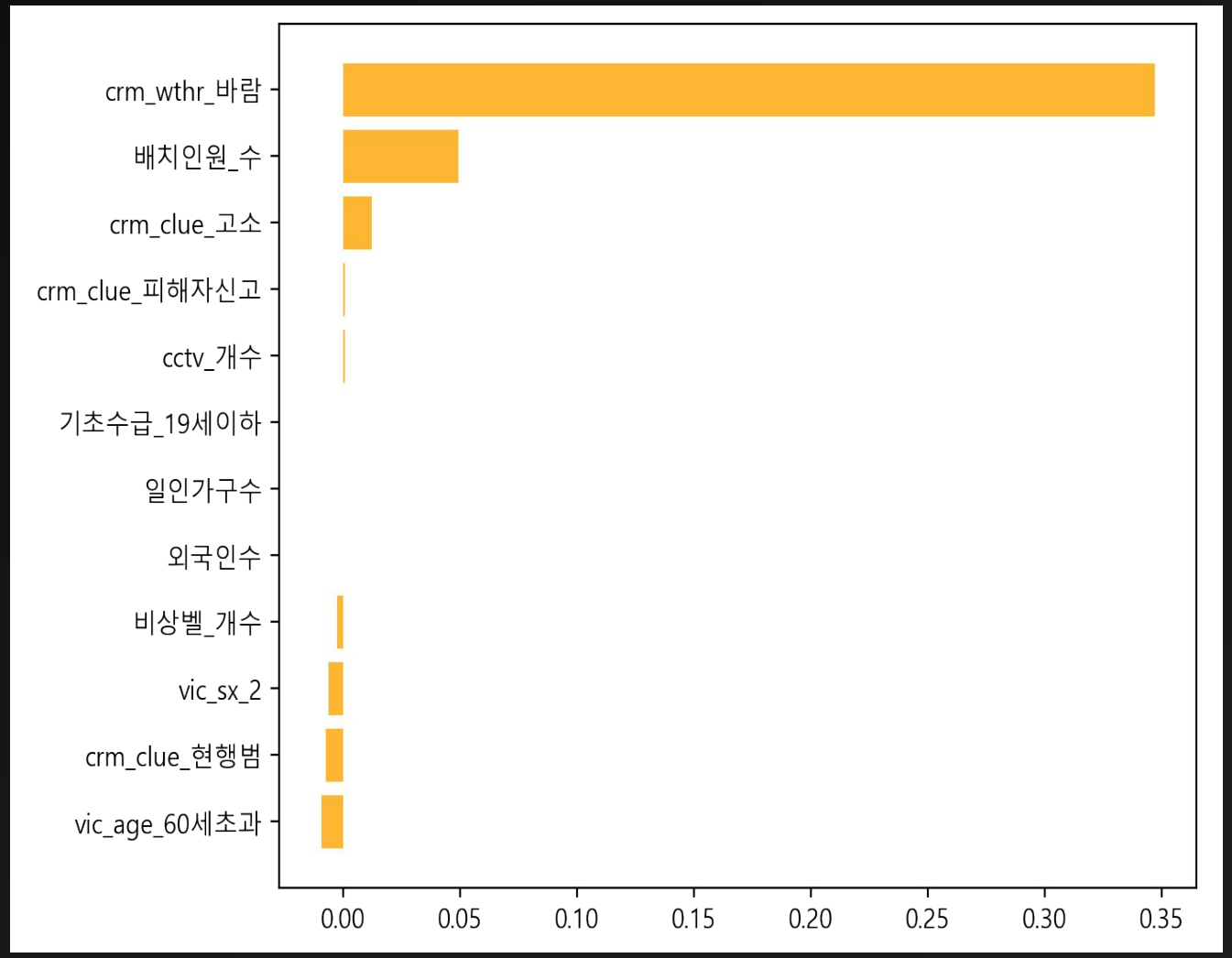
성능결과지표: MAE

검증방법 : K-fold교차검정 (n_splits = 5, random_state = 0)

안전도	스케일링	모델링(random_state = 0)	MAE(kfold교차검정의 평균값)
절도폭력	non-scaling	ridge(alpha = 10)	1.5
강도살인	robust-scaling	svr(kernel=linear)	1.81
교통사고	non-scaling	elasticnet(alpha=1)	1.525
법질서	minmax scaling	ridge(alpha=0.1)	1.89
전반적	non-scaling	ridge(alpha=100)	1.35

모델링 결과해석

절도폭력안전도



+ 양의영향

1.절도폭력 범죄발생시 **바람**이 많이 불었던 날씨의 건수가 **높을수록** 주거지역의 시민들의 **체감안전도가 높다.**

하지만, 상식적인 선에서 해석하는데 무리가 있는 결과라고 판단된다

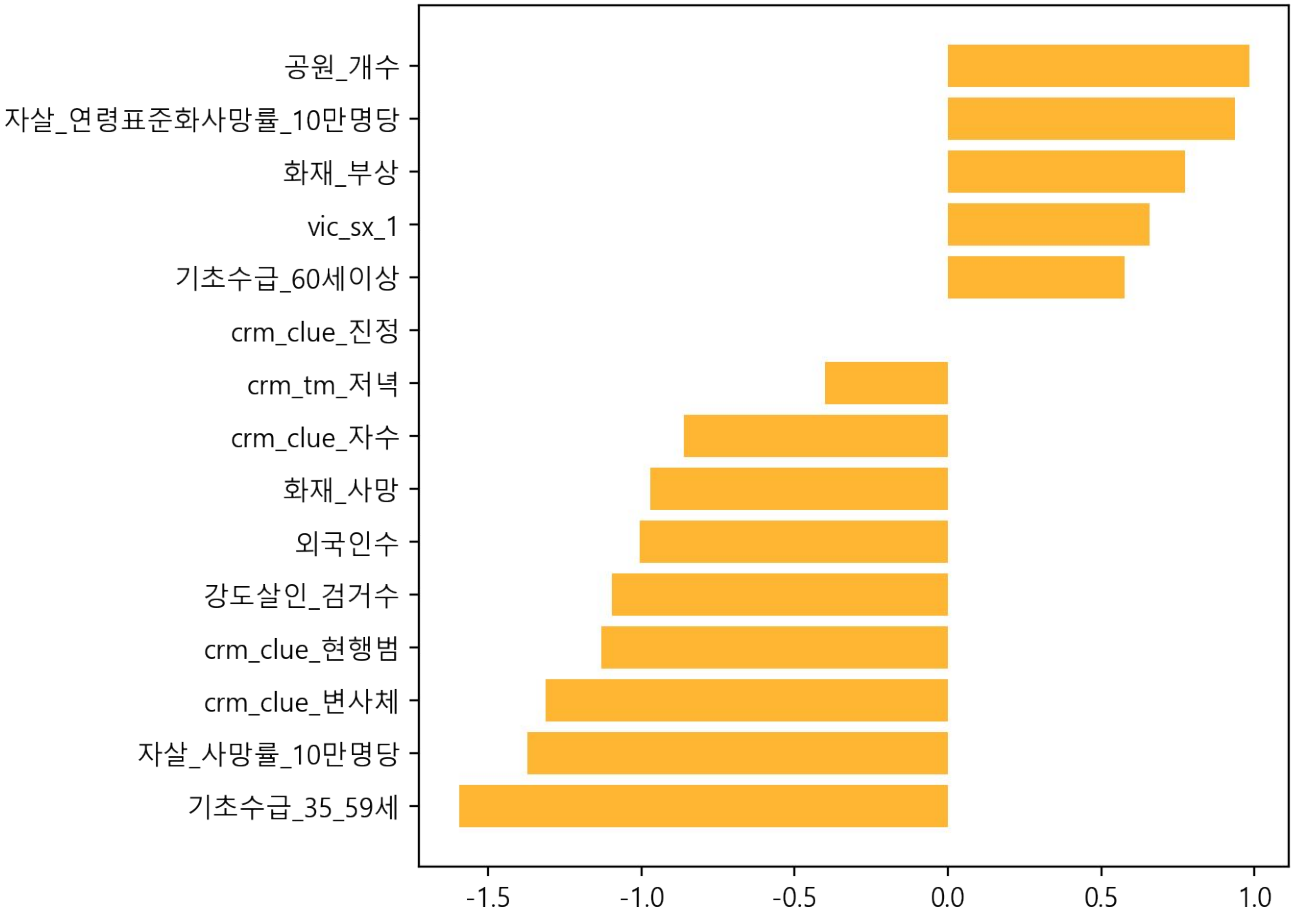
2.**지구대 별 배치인원수가 많을 수록 체감안전도가 높음.** 시민들의 절도폭력 안전도에 대한 긍정적인 영향을 미치는 변수로 고려된다

— 음의영향

1.**60세 이상 피해자의 수가 많을수록 절도폭력 체감안전도는 낮다**. 체감안전도 상승을 위해서는 절도폭력 고령 피해자에 대한 치안활동 개선점을 찾아야 할 것으로 보인다

모델링 결과해석

강도살인안전도



+ 양의영향

1.공원의 개수가 많을수록, 강도살인에 대한 체감안전도는 높다

— 음의영향

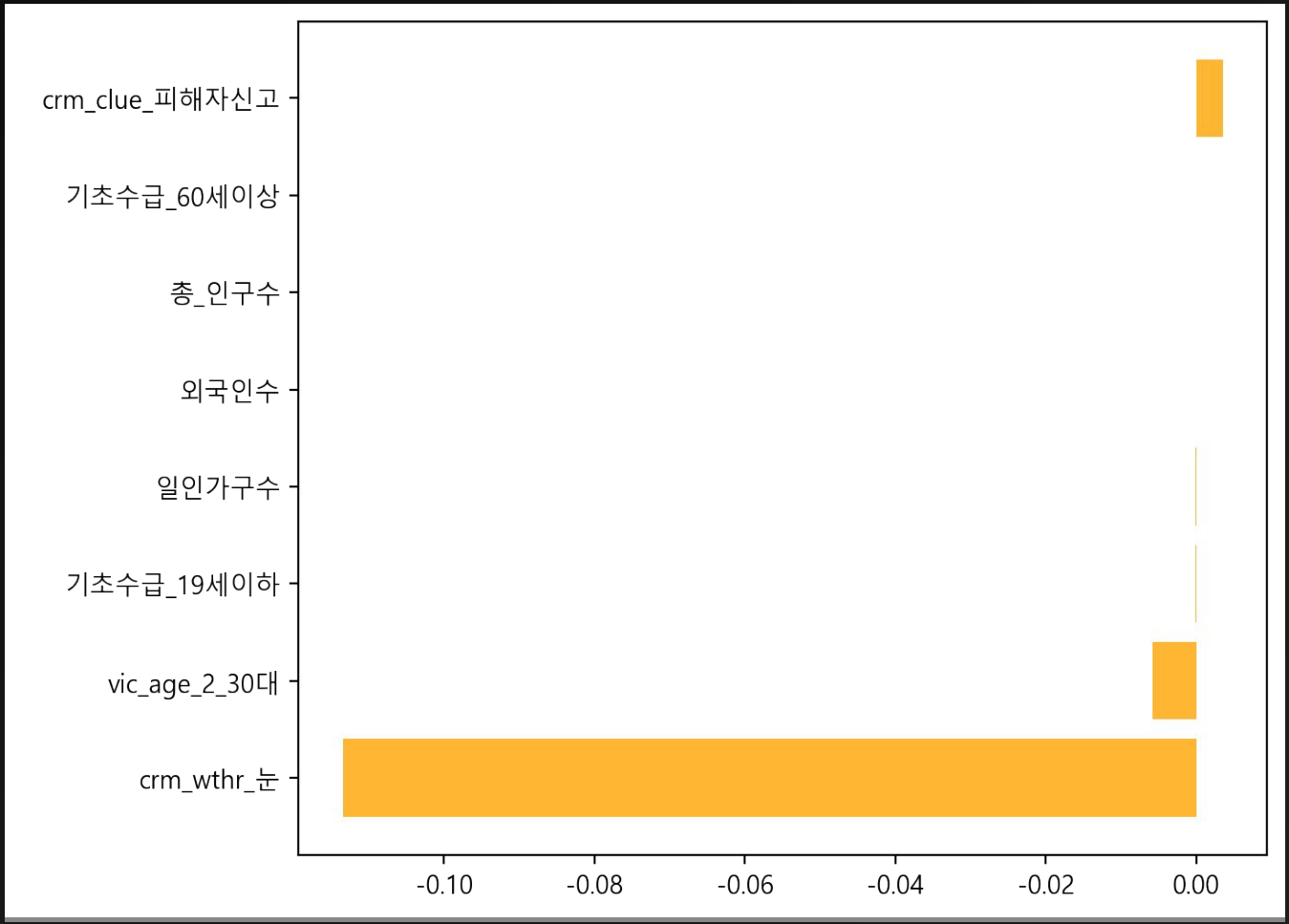
1.35~59세의 기초수급자 수가 많을수록 강도살인에 대한 체감안전도는 낮다. 또한, 10만명당 자살_사망률이 높을수록 체감안전도는 낮다.

두가지 음의 영향을 주는 변수를 종합적으로 해석한다면, 경제적, 환경적으로 열악한 상황에 있는 인구가 많을수록 시민들의 강도살인에 대한 두려움이 있음을 알 수 있다.

2.변사체 발견 수, 현행범, 강도살인의 검거수가 많을수록 강도살인 체감안전도는 낮다.

모델링 결과해석

교통사고안전도



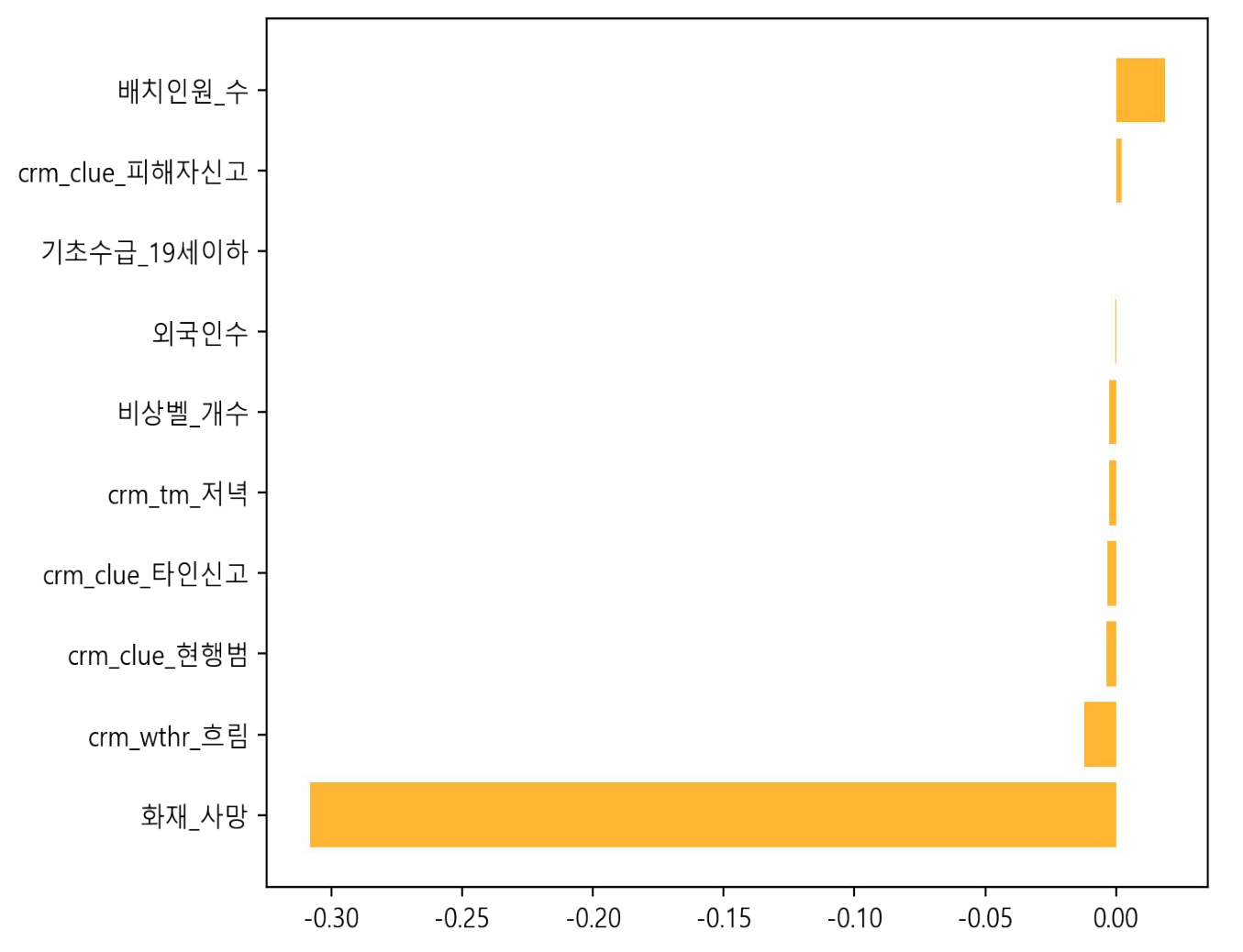
음의영향

1.**교통사고 범죄 발생 시, 눈이오는 날씨일수록 교통사고 체감안전도는 낮다.** 이는 눈으로인한 미끄러운 도로가 교통사고의 직접적인 영향을 주기때문에 시민들이 체감하는 안전도는 낮다고 해석될 수 있다.

Machine Learning

모델링 결과해석

전반적 안전도



+ 양의영향

1.**주거지역의 지구대 배치인원수가 많을수록 시민들의 전반적 체감안전도는 높다.**

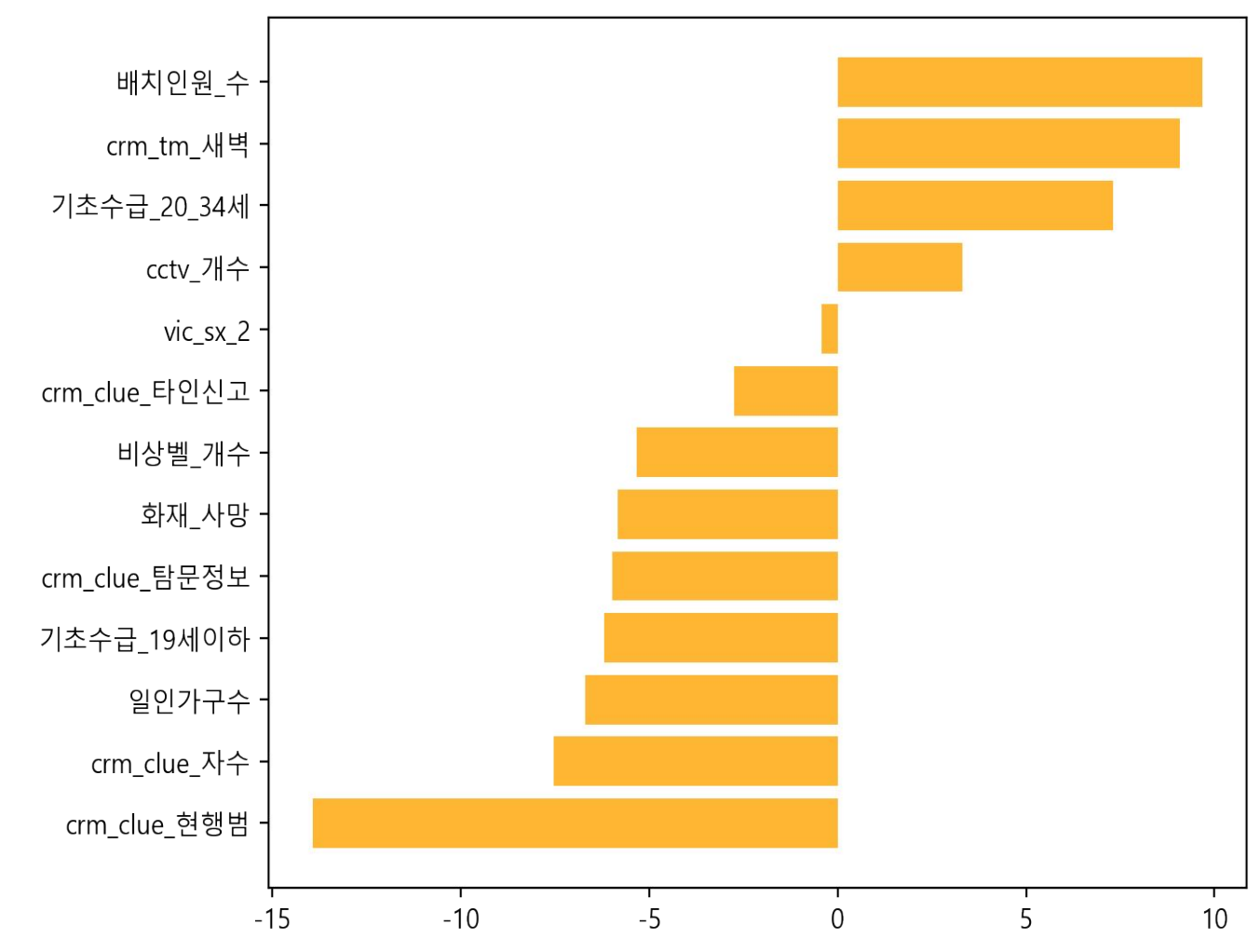
— 음의영향

1.**화재로 인한 사망수가 많을수록 전반적 체감안전도에는 음의 영향**을 미치는것으로 나타났다.

이는, 시민들이 전반적인 치안안전도에 있어 화재사건에 민감하다는 것으로 해석될 수 있다. 따라서, 전반적 안전도 제고를 위해서는 화재예방에 중점을 두어야 할 것으로 보인다.

모델링 결과해석

법질서안전도



양의영향

1.**배치인원수가 많을 수록, 범죄발생 시간대가 새벽일수록 cctv수가 많을수록 법질서 안전도에 대한 체감안전도는 높다.** 특히, 배치인원수, cctv수는 법질서 체감안전도에 긍정적인 영향을 주는 변수이므로, 법질서 안전도 제고를 위한 대안으로 고려된다.

음의영향

1.**현행범 자수, 1인가구수, 화재로 인한 사망건수 가 많을수록 법질서 체감 안전도에 음의 영향** 을 미친다.

법질서 체감안전도 개선을 위해서는 1인가구, 화재예방에 대한 주의깊은 관찰이 필요할 것으로 보인다

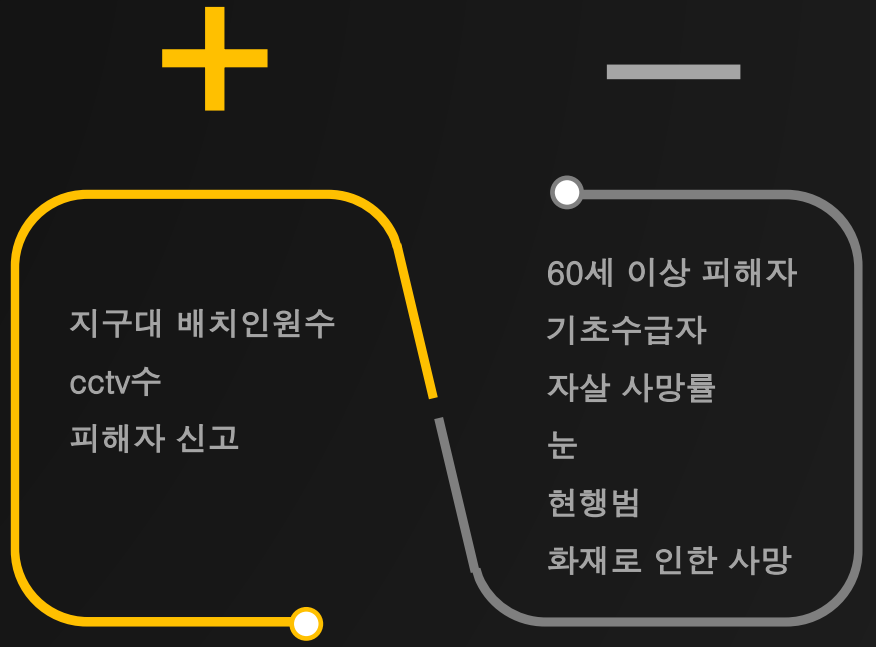
Machine Learning

체감안전도 예측-러닝머신러닝팀

Machine Learning

지구대 배치 인원

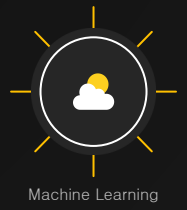
3개의 안전도에서 지구대 배치인원수가 양의 영향으로 나타났다. 이는, 시민들이 치안에 있어 **지구대 배치인원수가 많을수록 안정감**을 가지는 것으로 해석된다. 따라서, 치안안전도 제고를 위해서는 다(多),합리(合理)적인 지구대 배치인원수를 고려해야 할 것으로 보인다.



사회적 약자 인구

60세 이상 피해자, 기초수급자 등 사회적 약자에 해당하는 피해자 및 인구수가 체감안전도에 음의 영향을 준다. **사회적 약자 인구수가 많은** 관할서는 이를 위한 **치안안전활동에 더욱 중점**을 두어야 할 것으로 보인다. 또한, **화재사망에 대한 시민들의 민감도를 높으므로** 화재예방 활동 개선이 필요할 것으로 보인다.

- 결과
  각 안전도별 최고 성능을 모델로 20년 종합체감안전도를 예측한 결과, MAE 2.6으로 측정됨
- mae값이 안좋은 이유 및 한계점
  2020년도의 데이터셋이 주최측과 달라 mae값이 예상보다 차이가 큰 것으로 보인다.
  기존 머신러닝 모델만으로는 최고의 예측성능을 내는데 다소 아쉬운 것으로 판단된다.
- 개선방안
  딥러닝 모델로 성능을 최대치로 끌어올린다.



모델링 [딥러닝]

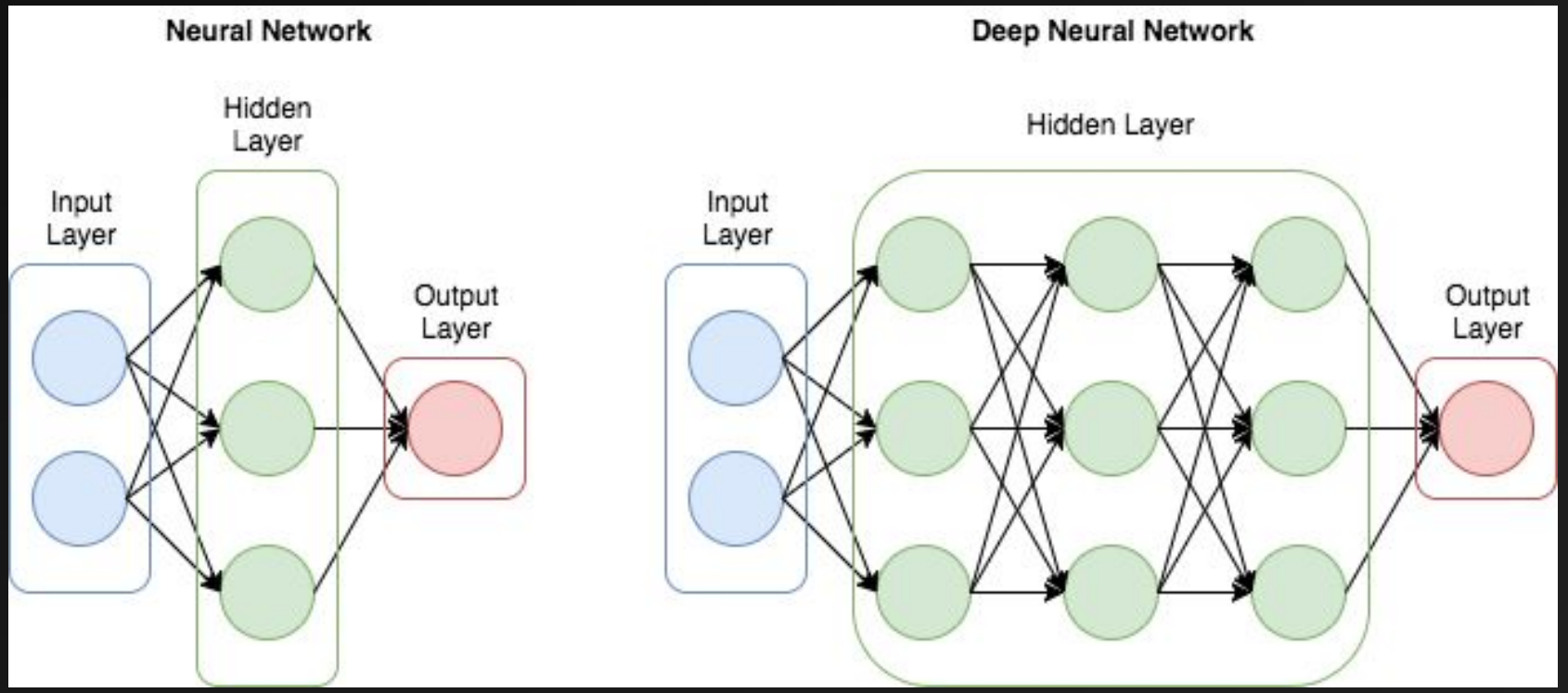
DNN 모델

DNN (Deep Neual Network) - 심층신경망

- 역전파 알고리즘(Backpropagation)으로 학습
- 경사하강법으로 에러(오차)를 최소화

DNN의 장점 (선택한 이유)

- 연속형, 범주형 상관없이 분석 가능
- 입력변수들간의 비선형조합 가능
- 다른 머신러닝에 비해 상대적으로 예측력 우수



이미지 출처:
https://ebbnflow.tistory.com/119

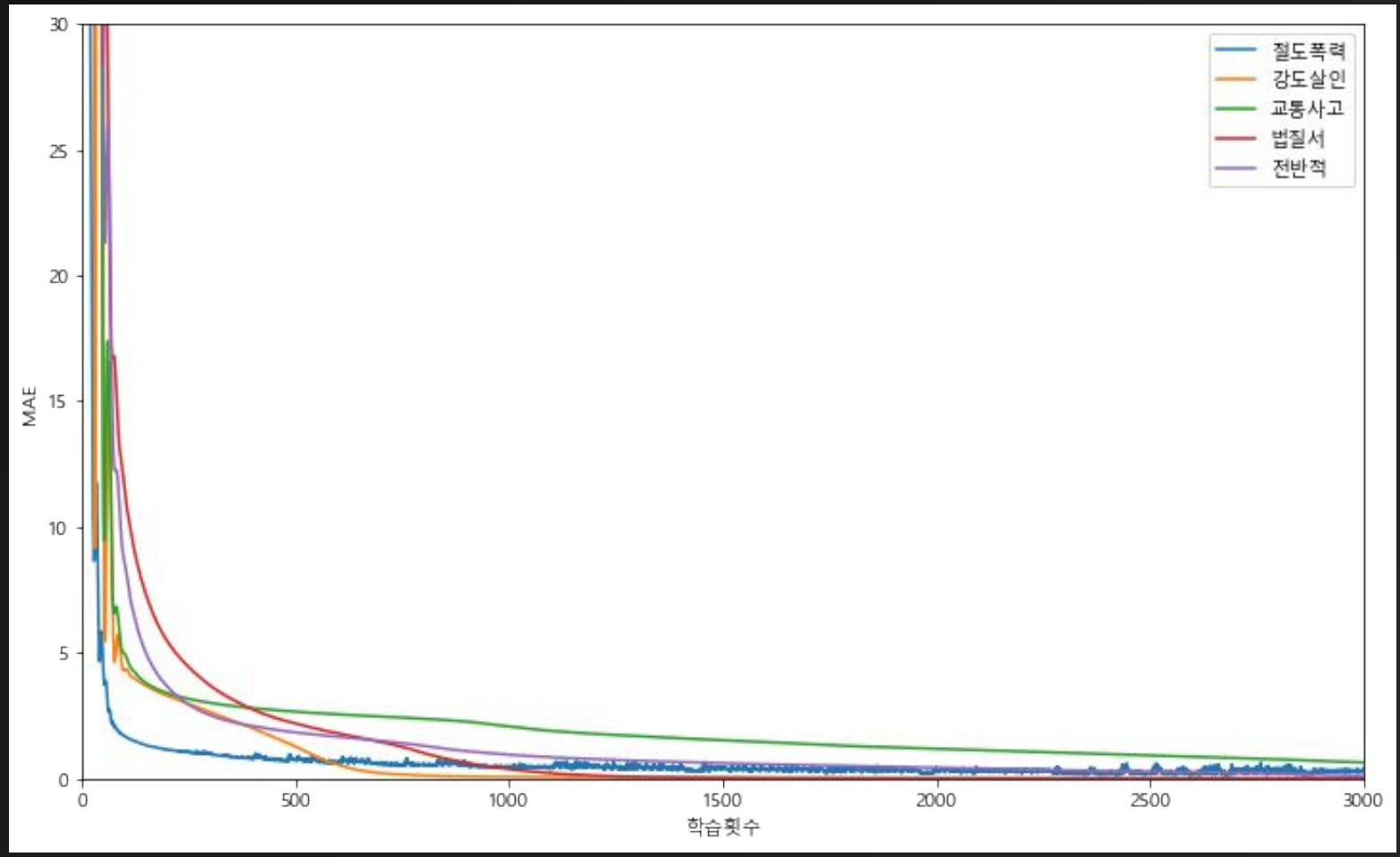
DNN 모델 튜닝

튜닝 값	비고
Xavier Initializer	적절한 초기가중치 설정을 통해 예측률 향상을 도모
relu	가장 많이 사용되는 활성화 함수로 예측률 향상에 도움
Learning rate : 0.001	학습률 조정
adam optimizer	기존의 Gradient Descent(경사하강법)보다 훨씬 좋은 성능을 내는 알고리즘
hidden layers : 5	hidden layer를 5개, 계층간 입출력 개수 512개로 wide하게 늘려 예측률 향상을 도모.
drop out : True	overfitting 방지

모델링 결과

강도살인안전도를 제외한 나머지 안전도에서는 **robust-scaling이 좋은 성능** 을 냈다
딥러닝 결과 학습횟수가 늘어남에 따라 **0.n점대로 수렴** 하는 것을 볼 수 있다.

안전도	스케일링
절도폭력	robust-scaling
강도살인	standard-scaling
교통사고	robust-scaling
법질서	robust-scaling
전반적	robust-scaling



예측 결과 한계점

결과

각 안전도의 학습횟수를 10만번으로 지정후, 20년 종합체감안전도 예측한 결과 기존의 mae 2.6에서 딥러닝을 통해 1.50으로 예측성능을 높일 수 있었다

- 한계점

학습횟수가 많기 때문에 시간이 오래 걸린다

딥러닝 모델은 변수영향도를 분석하기 어렵다는 점에서 한계점이 있다

모델링 과정에서는 0.002의 매우 적은 오차값을 가졌지만 머신러닝 예측과 마찬가지로 20년도 데이터셋 구축에 문제가 있는 것으로 보인다. 주최측과 같은 20년도의 데이터셋을 통해 예측한다면 0.n점대의 오차값을 도출할 것으로 보인다

데이터 출처
경찰범죄통계사이트 https://www.police.go.kr/www/open/publice/publice03_2020.jsp
compas치안체감안전도데이터 https://compas.lh.or.kr/subj/competition/data?subjNo=SBJ_2107_004
국가통계포털 KOSIS https://kosis.kr/index/index.do

체감안전도 예측

COMPAS 러닝 머신 러닝 팀

시민이 공감하는 치안 체감안전도 예측