체감안전도 예측

에이콘 아카데미 5조 발표

양정* 김일* 이현* 박병* 전우*

Machine Learning

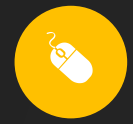
Perceived Safety Contents

p 1

Contents

전처리 핵심

결측치 처리 구간화 위치 데이터 이용 후 변주형 변수 (One hot-encoding)



모델링 핵심

변수선택 시 차원의 저주 머신러닝 MAE, kfold

Perceived Safety Subject Outline

Machine Learning

Perceived Safety Subject Outline

Machine Learning

분석배경

Safety

범죄 재난 등으로부터
안전한 국민의 삶을 위해

Policy

다양한 치안정책 사업에
도움을 위해

Feeling

국민이 체감할 수 있는
성과로 이어질 수 있도록

해결 목표

2019
시민이 공감하는 치안 체감안전도

Factors

다양한 요인을 분석하여
실시간 안전도 예측

Improvement

경찰이 가진 자원 주민들이 가진 자원을 활용하여 안전도 개선

Perceived Safety Subject Outline

Machine Learning

41개 관할서

서울과 경기, 경남 일부 지역 41개 관할서 별 2019년도 체감안전도

2017~19년도 데이터

112신고, 범죄발생원표, 범죄검거원표, 보안등, cctv, 교통사고, 지구대별인원현황, 화재발생통계, 성연령별인구분포, 1인가구수, 외국인 인구수, 기초수급자현황, 최종학력통계, 공원현황, 유흥업소 및 주점 현황, 비상벨현황 등

범죄명 분류를 위해 범죄 발생원표를 이용

● ○ ○ ○ ○ ○ ○ ○ ○ ○ " 현재 주거 지역은 절도, 폭력 등과 같은 범죄로부터 얼마나 안전한가"

● ● ○ ○ ○ ○ ○ ○ ○ ○ " 강도 살인 등과 같은 범죄로부터 얼마나 안전한가"

● ● ● ○ ○ ○ ○ ○ ○ ○ " 교통 사고로부터 얼마나 안전한가"

● ● ● ● ○ ○ ○ ○ ○ ○ " 기초질서, 집회시위질서 등 법질서적으로 얼마나 안전한가"

● ● ● ● ● ○ ○ ○ ○ ○ " 모든 항목을 통합하여 얼마나 안전한가"

Perceived Safety Data Preprocessing

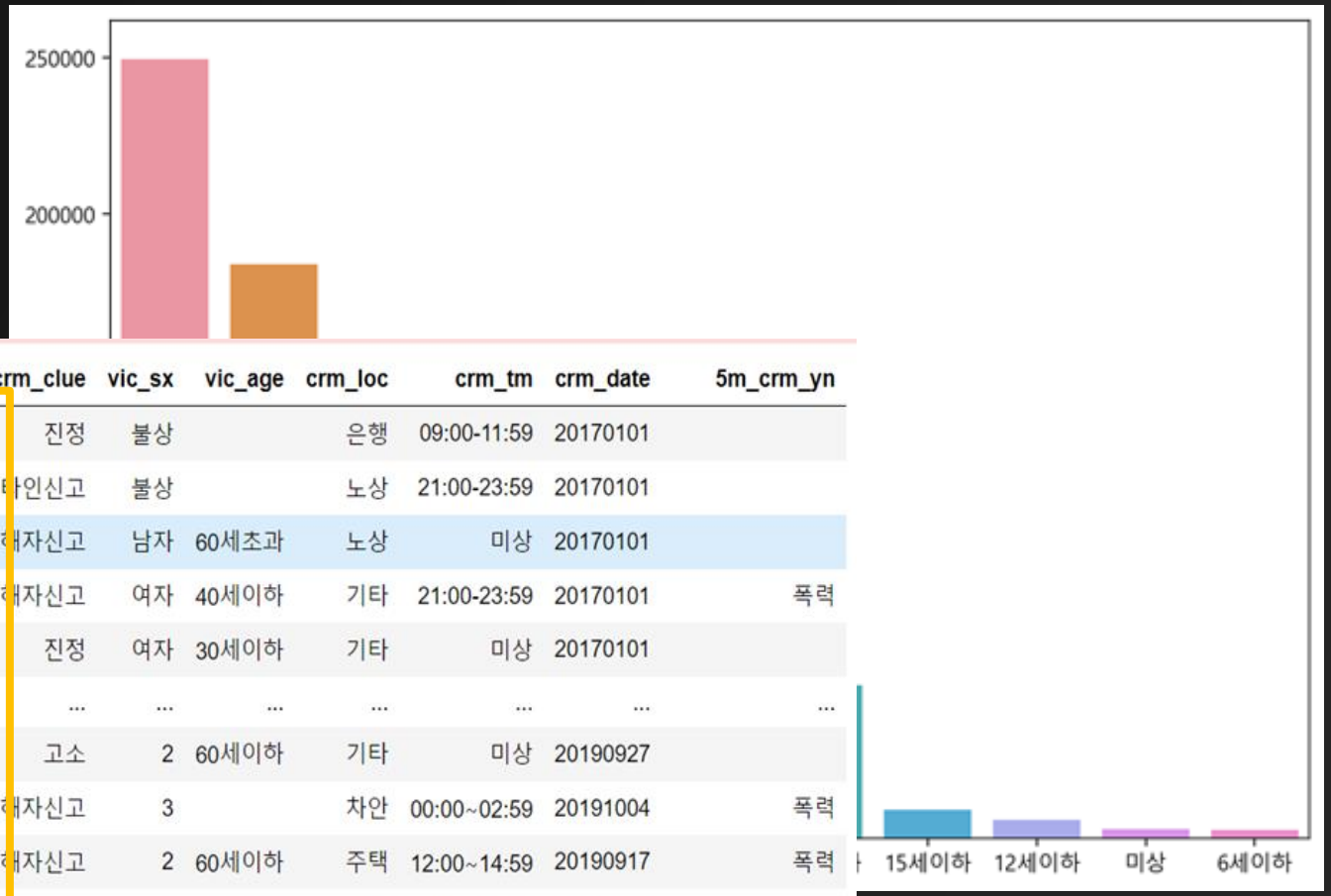
Machine Learning

결측치 처리

피해자 연령대

결측치 비율이 높아 제거하기 어려움

범주형 변수이므로 '미상'으로 표시



	jur_stn	crm	crm_wthr	crm_clue	vic_sx	vic_age	crm_loc	crm_tm	crm_date	5m_crm_yn
0	서울수서경찰서	위조외국통화행사	미상	진정	불상		은행	09:00-11:59	20170101	
1	서울영등포경찰서	도로교통법위반	맑음	타인신고	불상		노상	21:00-23:59	20170101	
2	서울양천경찰서	209015100	미상	피해자신고	남자	60세초과	노상	미상	20170101	
3	서울서초경찰서	폭행	미상	피해자신고	여자	40세이하	기타	21:00-23:59	20170101	폭력
4	서울동대문경찰서	사기	미상	진정	여자	30세이하	기타	미상	20170101	
...	...	...	...	...	...	...	...	...	...	...
1068235	경남진해경찰서	사기	미상	고소	2	60세이하	기타	미상	20190927	
1068236	경남진해경찰서	폭행	맑음	피해자신고	3		차안	00:00~02:59	20191004	폭력
1068237	경남마산동부경찰서	재물손괴	미상	피해자신고	2	60세이하	주택	12:00~14:59	20190917	폭력
1068238	경남마산중부경찰서	사기	미상	진정	1	50세이하	기타	12:00~14:59	20190914	
1068239	경남마산동부경찰서	강제추행	미상	고소	1	30세이하	기타	21:00~23:59	20190516	강간 및 강제추행

1068240 rows × 10 columns

112 신고 파일

두 변수 결측치 존재

참고할 변수가 없어 결측치 제거,
신고 내용 변수만 결측치 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

```
print(len(call_songpa_18.value_counts('inc_info')))
print(call_songpa_18.value_counts('inc_info').head())
```

```
48
inc_info
기타형사범     16027
보호조치        9691
상담문의        8463
시비            6223
교통사고        5993
dtype: int64
```

```
(call_songpa_18.value_counts('inc_info')₩
 / sum(call_songpa_18.value_counts('inc_info').values))₩
.head(10)
```

```
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
```

신고 내용 확인 가능 변수

모집단 가정

48개의 경우의 수를 **이산형 확률분포**로 가정

빈도 수/전체로 만들어 inc_info를 n=48인 **다항분포**로 정의

신고 내용 확인 불가 변수

모집단의 표본

다항분포를 따르는 결과라 예상

```
call_songpa_18_prob = call_songpa_18.value_counts('inc_info') ₩
/ sum(call_songpa_18.value_counts('inc_info').values)
print(np.random.multinomial(len(songpa_18_na),
                    pvals=call_songpa_18_prob.values, size=30))
songpa_18_replace=[]
for i in range(len(call_songpa_18.value_counts('inc_info').index)):
    a = np.mean(multi_songpa_18[:,i])
    songpa_18_replace.append(a)
print(songpa_18_replace[0:5])
```

```
[[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]
```

대체값 추론

Numpy의 다항분포함수

30회의 랜덤추출을 반복으로
평균값을 구해 분포 추론

```
songpa_18_replace = pd.Series(songpa_18_replace, index=call_songpa_18.value_counts('inc_info').keys())
songpa_18_replace.head(10)
```

```
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
```

Index로 결측치 대체

1인가구

연도 부재

인구 증가가 일정한 선형으로 이루어진다고 가정

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

20년과 15년 인구의 차를 5로 나누어 17,18,19년도 인구를 계산

```
def solo_calc(df):
    increase = (df.sum()[1] - df.sum()[0]) / 5
    solo_2016 = np.round(df.sum()[0]+increase)
    solo_2017 = np.round(solo_2016+increase)
    solo_2018 = np.round(solo_2017+increase)
    solo_2019 = np.round(solo_2018+increase)
    solo_dict = {'2017' : solo_2017, '2018' : solo_2018, '2019' : solo_2019}
    return pd.Series(solo_dict)
```

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

결과

구간화

시간대,나이 변경

1.3시간 간격으로 값이 측정되어 변경

crm_tm
09:00-11:59
21:00-23:59
미상
21:00-23:59
미상
...
미상
00:00~02:59
12:00~14:59
12:00~14:59
21:00~23:59

crm_tm_미상	crm_tm_새벽	crm_tm_오전	crm_tm_오후	crm_tm_저녁
0	5	0	2	6
0	3	0	4	3
0	4	0	4	4
0	2	0	2	3
1	3	2	1	2
0	2	2	6	3
1	0	0	1	0
1	2	1	0	6
0	2	1	2	1
0	4	4	6	5

age
4
45
46
47
48
...
39
40
41
42
43

	19세 이하	20 ~ 34세	35 ~ 59세	60세 이상
은평구	44183	22300	72209	111644
강서구	41381	25216	88163	152898
성북구	30625	13009	48861	79450
세종특별자치시	19178	5176	20354	27770
영등포구	13120	6719	35595	61550
용산구	10987	5343	27685	44773
수원시	48972	18582	77990	122389
동대문구	23646	10627	47479	85930
관악구	28427	20072	66336	100761
강북구	35548	16872	65633	103682
광진구	22892	12362	42650	52905
강동구	24777	12227	46896	68456
진주시	36274	11965	50900	64143
종로구	5516	2985	16072	28596

위치 데이터 처리

관할서 파생변수 생성

데이터 전처리

위도경도 정보를 통해 관할서 지정

	address	securitylight_cnt	lon	lat
0	세종특별자치시 금남면 감성리 64-2	1	127.287690	36.443467
1	세종특별자치시 금남면 감성리 267	1	127.288812	36.444181
2	세종특별자치시 금남면 감성리 40-1	1	127.289575	36.444711
3	세종특별자치시 금남면 감성리 26	1	127.290071	36.444455
4	세종특별자치시 금남면 감성리 267	1	127.290002	36.444188
...	...	...	...	...
229400	서울특별시 중랑구 상봉동 19-2	1	127.092434	37.602786
229401	서울특별시 중랑구 상봉동 19-44	1	127.092721	37.602497
229402	서울특별시 중랑구 상봉동 19-25	1	127.092258	37.602374
229403	서울특별시 중랑구 상봉동 495-4	1	127.094617	37.602357
229404	서울특별시 중랑구 묵동 3-5	1	127.083960	37.618769

229405 rows × 4 columns

위치 데이터 처리

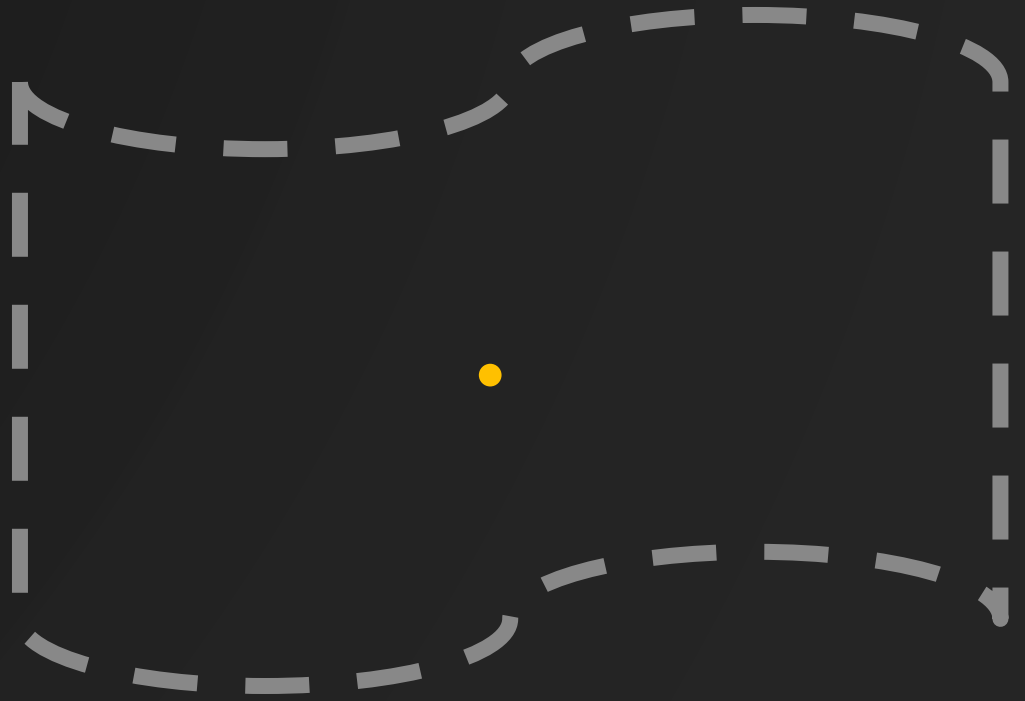
관할서 파생변수 생성

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



위치 데이터 처리

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

위치 데이터 처리

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
...	...	...	...	...	...

범주형 변수

One hot-encoding

	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

Perceived Safety Modelling

Machine Learning

Perceived Safety Modelling

Machine Learning

변수선택

차원의 저주

차원, 변수의 증가에 따라 성능이 저하됨

절도폭력안전도 변수	69개	123개
강도살인안전도 변수	62개	123개
교통사고안전도 변수	129개	123개

Coefficients: (3 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.483e+01	4.665e+00	16.042	< 2e-16	***
화재_인명피해_계	NA	NA	NA	NA	
화재_부동산피해_천원	-1.486e-06	7.519e-07	-1.976	0.05265	.
화재_동산피해_천원	-1.335e-06	6.451e-07	-2.069	0.04274	*
화재_재산피해_천원	NA	NA	NA	NA	
화재_재산피해_건당천원	2.855e-04	1.376e-04	2.074	0.04221	*
cctv_개수	3.280e-04	3.134e-04	1.046	0.29941	
기초수급_60세이상	5.967e-05	5.667e-05	1.053	0.29651	
기초수급_총합	NA	NA	NA	NA	

변수선택

후진제거법

다중공선성이 낮아지지 않고, 상관계수가 높은 변수 많음

모든 변수를 아우르지 못함

Dep. Variable:	score_절폭	R-squared:	0.638
Model:	OLS	Adj. R-squared:	0.595
Method:	Least Squares	F-statistic:	14.78
Date:	Fri, 20 Aug 2021	Prob (F-statistic):	1.33e-18
Time:	18:25:13	Log-Likelihood:	-258.41
No. Observations:	123	AIC:	544.8
Df Residuals:	109	BIC:	584.2
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	76.4938	0.740	103.351	0.000	75.027	77.961
crm_wthr_바람	0.2761	0.140	1.969	0.051	-0.002	0.554
crm_clue_타인신고	-0.0167	0.008	-2.024	0.045	-0.033	-0.000
crm_clue_현행범	-0.0053	0.001	-4.556	0.000	-0.008	-0.003
vic_age_60세초과	-0.0122	0.003	-4.686	0.000	-0.017	-0.007
화재_사망	-0.1326	0.102	-1.301	0.196	-0.334	0.069
cctv_개수	0.0005	0.000	2.546	0.012	0.000	0.001
배치인원_수	0.0495	0.007	7.165	0.000	0.036	0.063
비상벨_개수	-0.0024	0.001	-3.783	0.000	-0.004	-0.001
외국인수	-8.479e-05	2.24e-05	-3.791	0.000	-0.000	-4.05e-05
자살건수	-0.0554	0.020	-2.799	0.006	-0.095	-0.016
vic_sx_1	-0.0012	0.001	-1.245	0.216	-0.003	0.001
화재_부상	0.0472	0.024	1.936	0.055	-0.001	0.096
일인가구수	-6.455e-05	2.08e-05	-3.099	0.002	-0.000	-2.33e-05

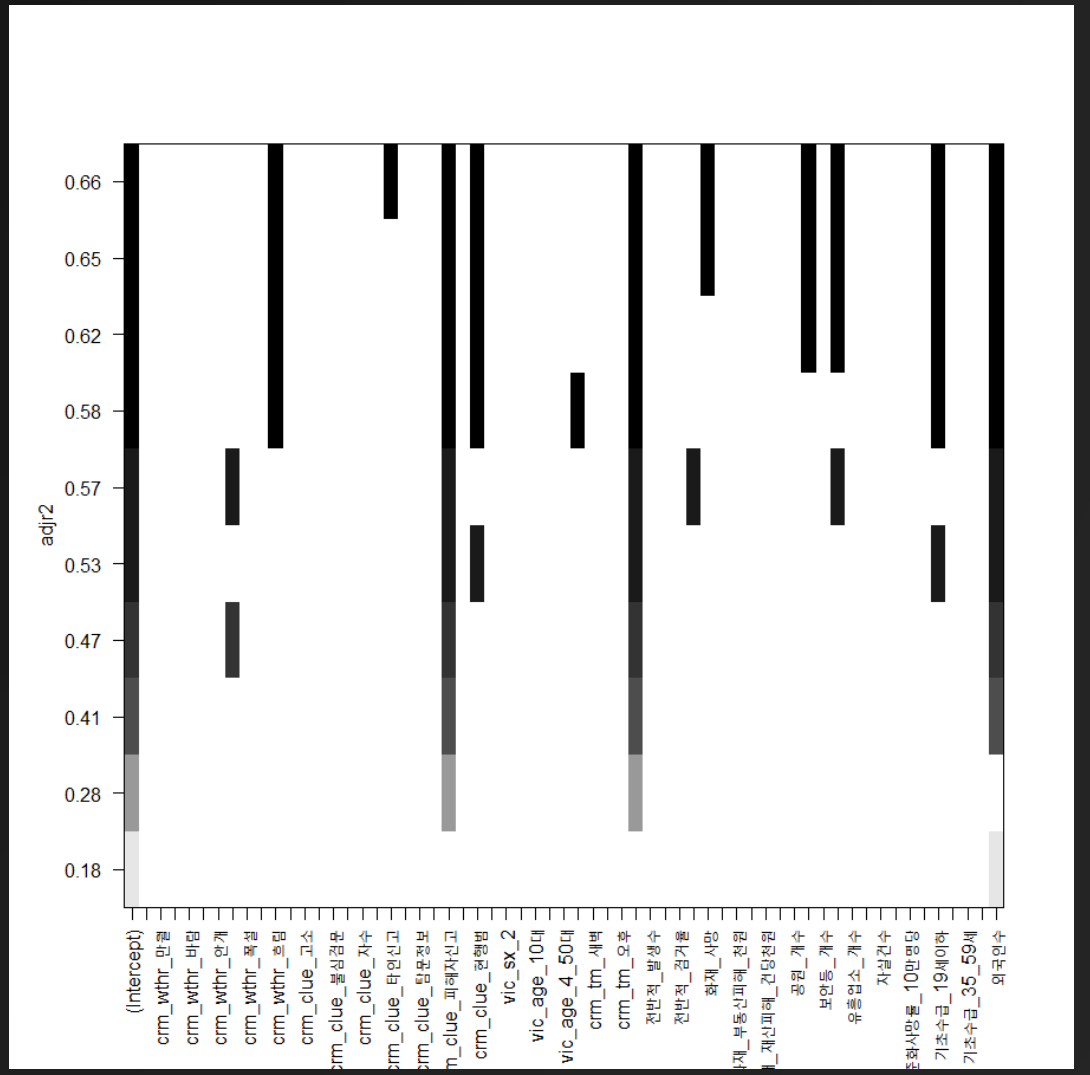
	VIF Factor	features
0	inf	const
1	inf	crm_wthr_바람
2	inf	crm_clue_타인신고
3	inf	crm_clue_현행범
4	inf	vic_age_60세초과
5	inf	화재_사망
6	4503599627370496.00000	cctv_개수
7	inf	배치인원_수
8	900719925474099.25000	비상벨_개수
9	inf	외국인수
10	inf	자살건수
11	inf	vic_sx_1
12	inf	화재_부상
13	inf	일인가구수

변수축소

차원의 저주 해결

R의 All Subset regression
(부분집합회귀분석)

모든 변수 사용 가장 좋은 성능(결정 수정계수)을 내는 n개의 변수를 추출



Perceived Safety Modelling

Machine Learning

추출 변수

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

Perceived Safety Machine Learning

성능테스트

모델 리스트

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 결과

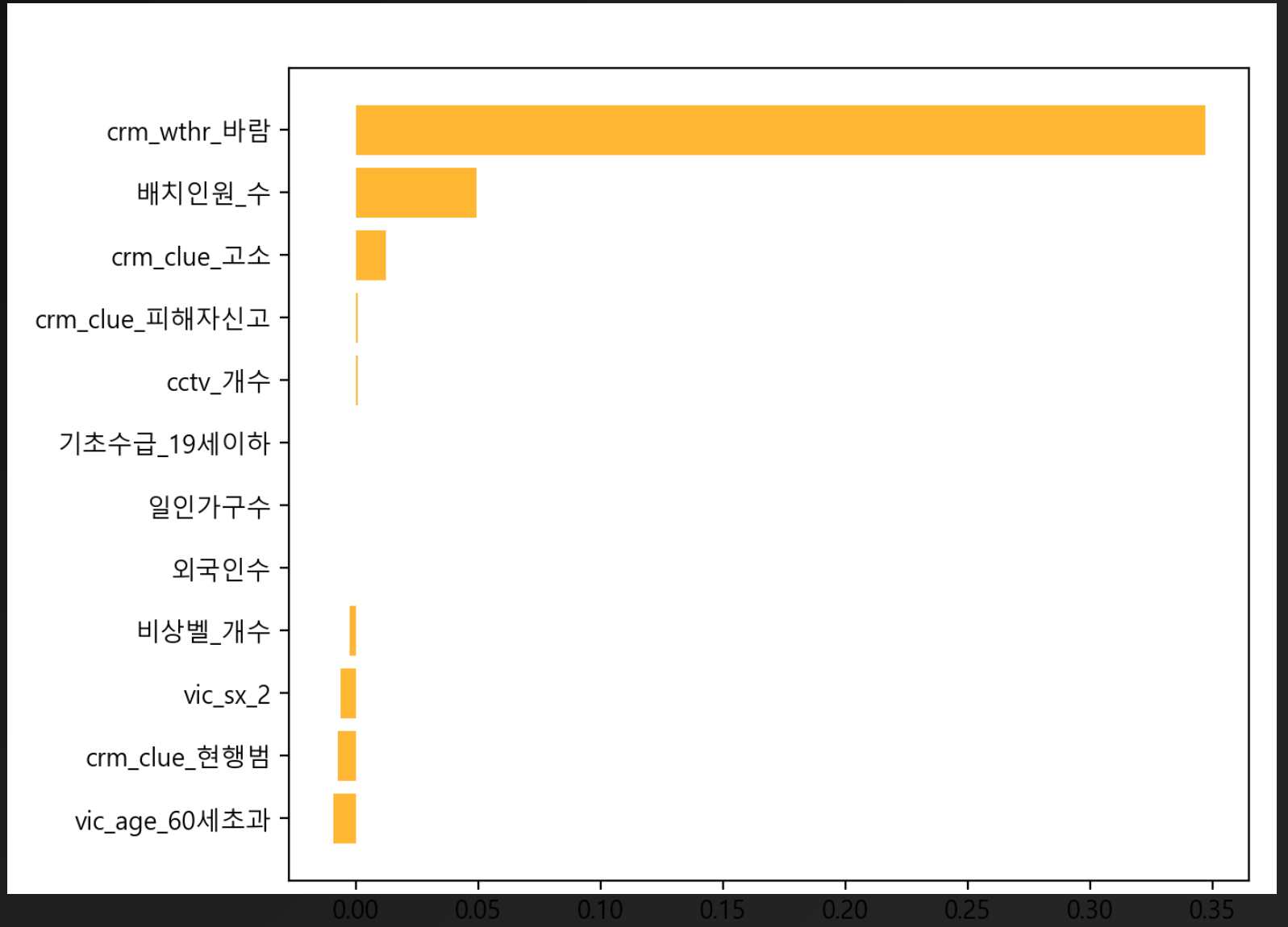
성능결과지표: MAE
검증방법: K-fold교차검정

안전도	스케일링	모델링	MAE
절도폭력	non-scaling	elasticnet(alpha=0.01)	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

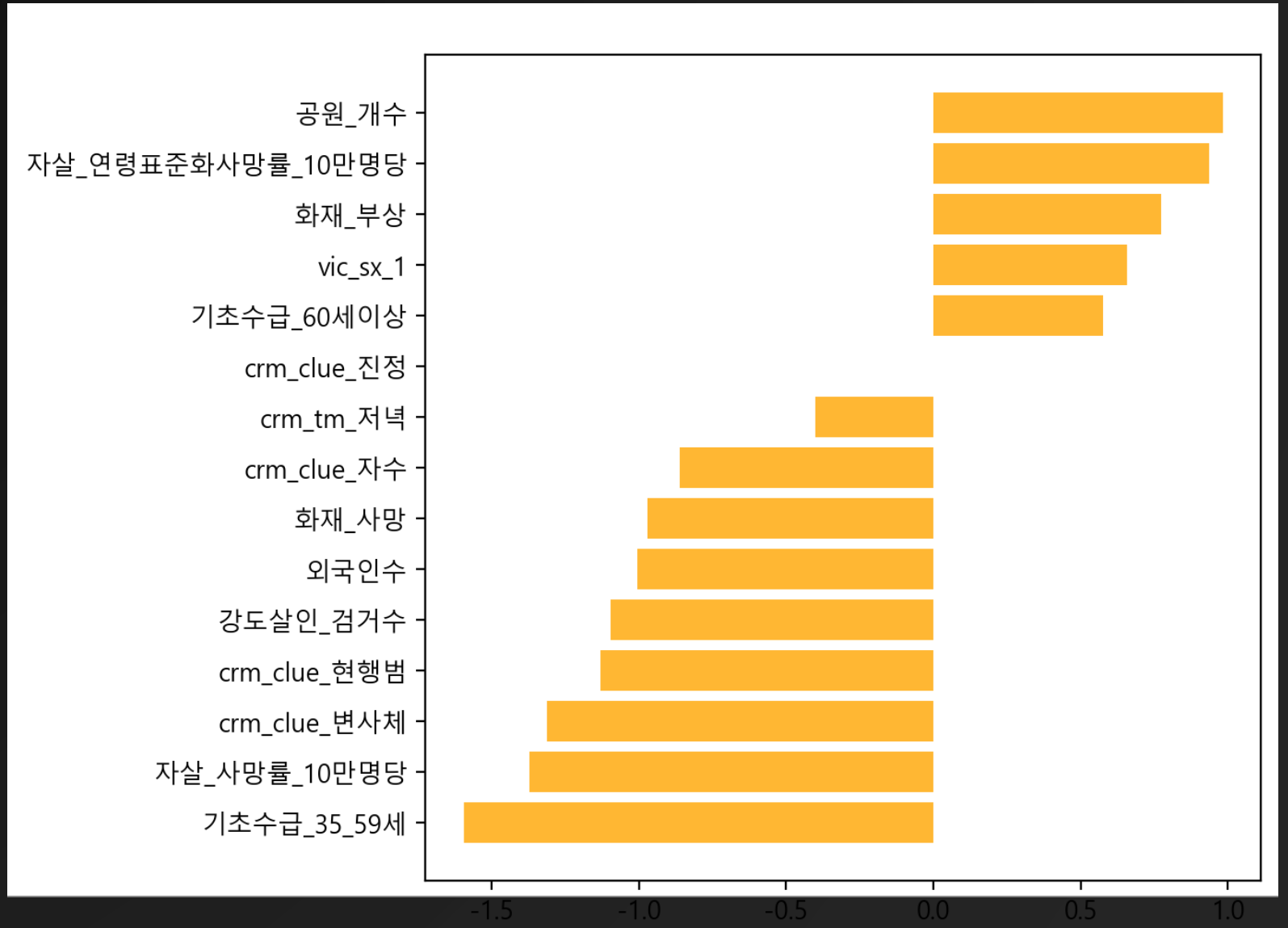
Perceived Safety Result Analysis

Machine Learning

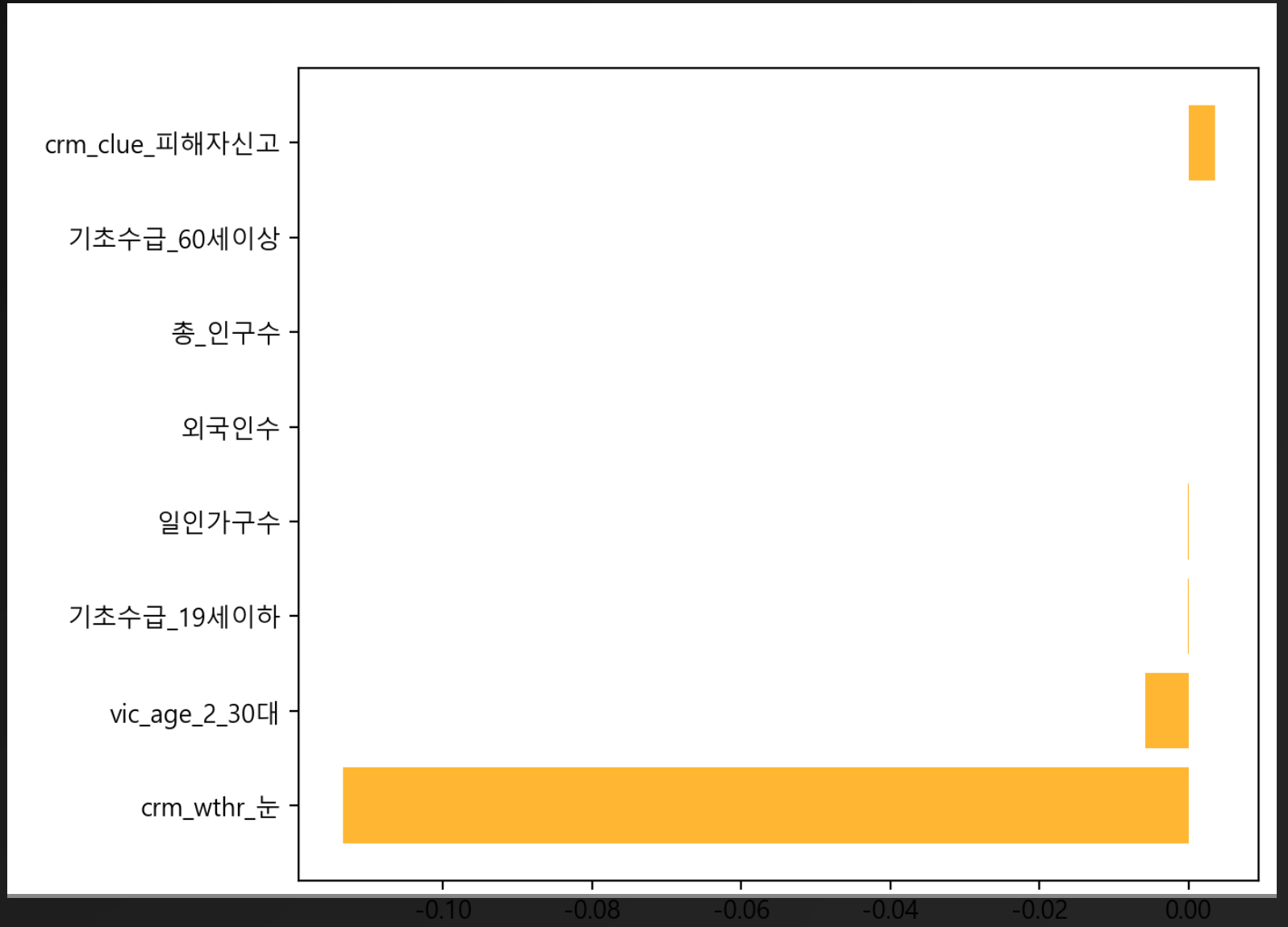
절도폭력안전도


crm_wthr_바람
배치인원_수
crm_clue_고소
crm_clue_피해자신고
cctv_개수
기초수급_19세이하
일인가구수
외국인수
비상벨_개수
vic_sx_2
crm_clue_현행범
vic_age_60세초과
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35

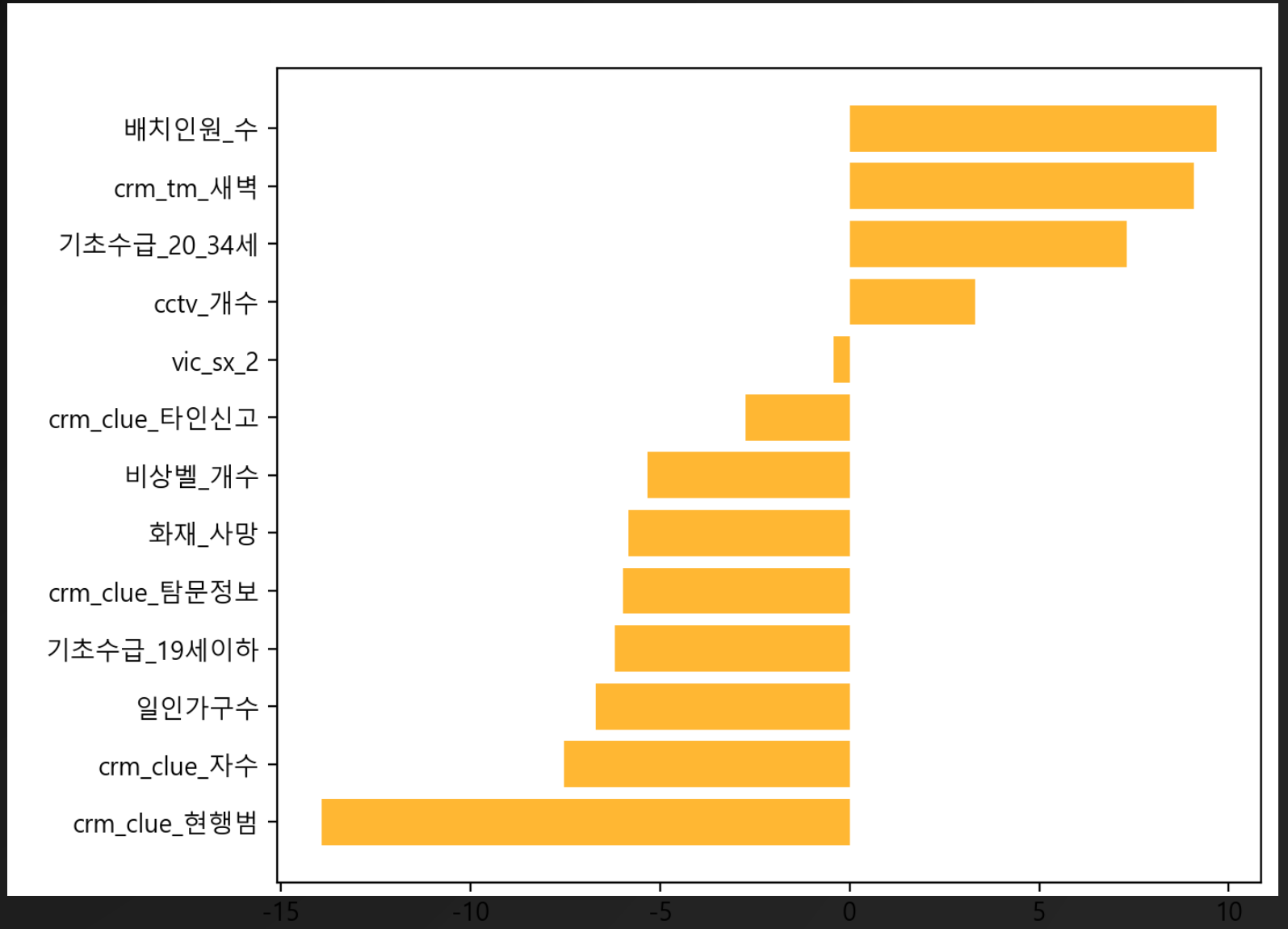
강도살인안전도


공원_개수
자살_연령표준화사망률_10만명당
화재_부상
vic_sx_1
기초수급_60세이상
crm_clue_진정
crm_tm_저녁
crm_clue_자수
화재_사망
외국인수
강도살인_검거수
crm_clue_현행범
crm_clue_변사체
자살_사망률_10만명당
기초수급_35_59세
-1.5
-1.0
-0.5
0.0
0.5
1.0

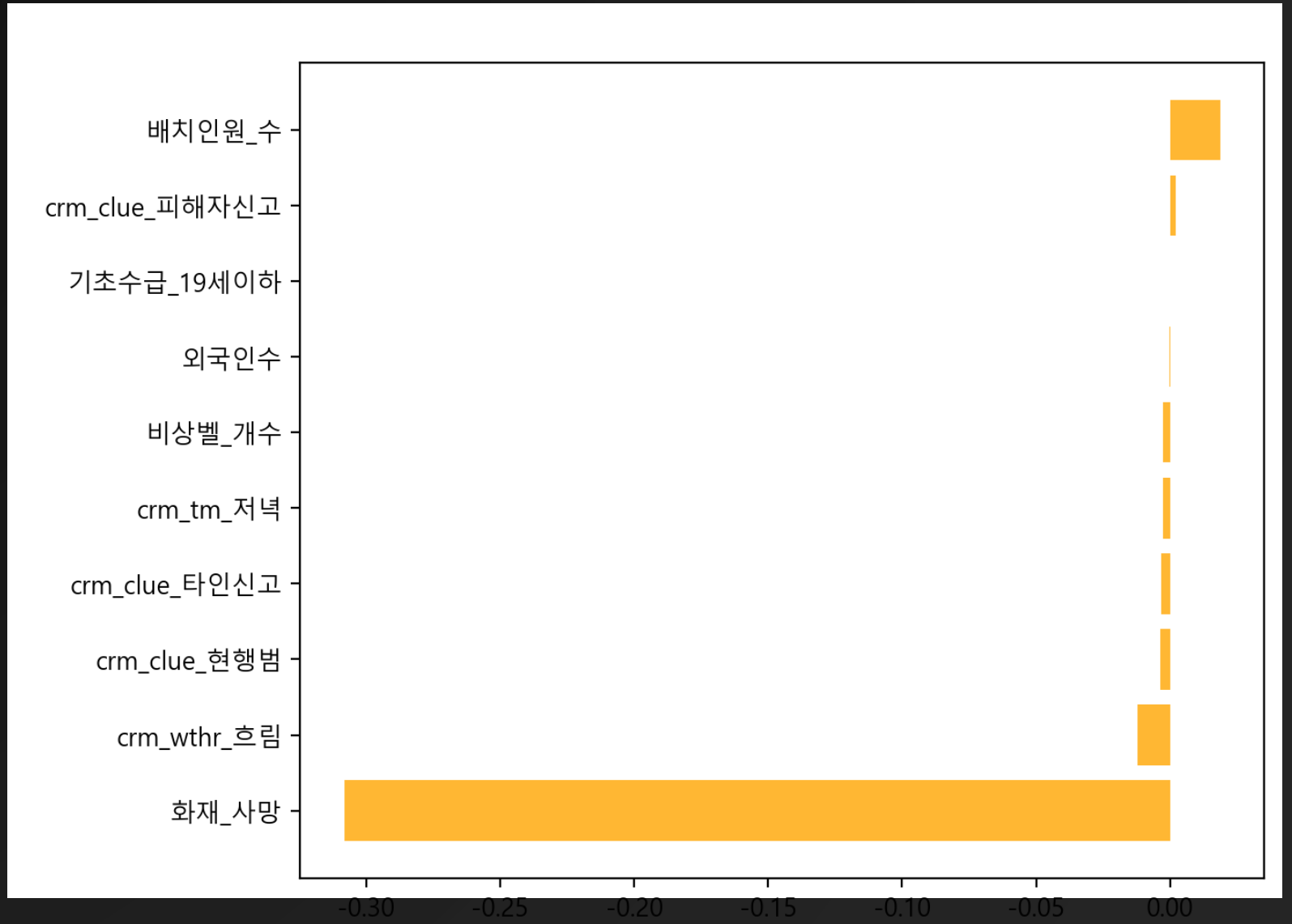
교통사고안전도


crm_clue_피해자신고
기초수급_60세이상
총_인구수
외국인수
일인가구수
기초수급_19세이하
vic_age_2_30대
crm_wthr_눈
-0.10
-0.08
-0.06
-0.04
-0.02
0.00

법질서안전도


배치인원_수
crm_tm_새벽
기초수급_20_34세
cctv_개수
vic_sx_2
crm_clue_타인신고
비상벨_개수
화재_사망
crm_clue_탐문정보
기초수급_19세이하
일인가구수
crm_clue_자수
crm_clue_현행범
-15
-10
-5
0
5
10

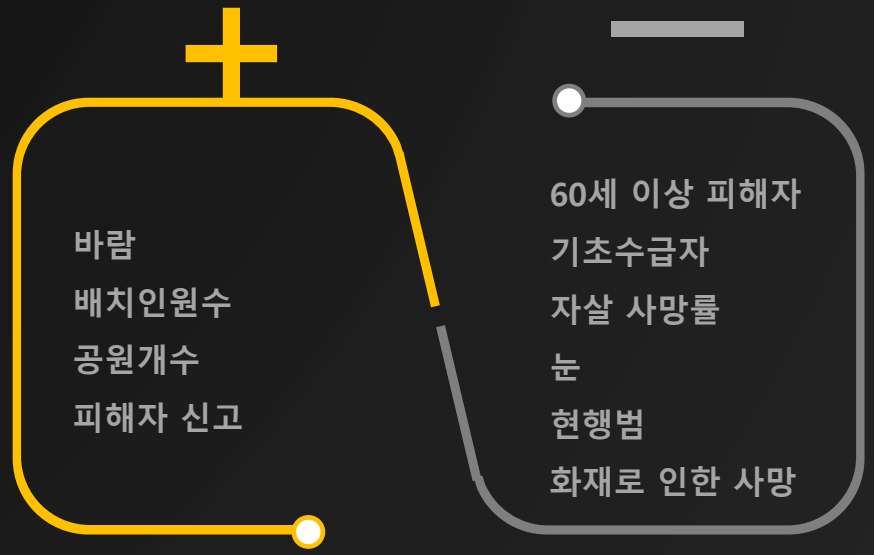
전반적 안전도


배치인원_수
crm_clue_피해자신고
기초수급_19세이하
외국인수
비상벨_개수
crm_tm_저녁
crm_clue_타인신고
crm_clue_현행범
crm_wthr_흐림
화재_사망
-0.30
-0.25
-0.20
-0.15
-0.10
-0.05
0.00

Perceived Safety Result Analysis

지구대 배치 인원

해당지역의 지구대 배치 인원이
높을 수록 체감안전도가 높다



사회적 약자 인구

사회적 약자에 해당하는 피해자 및 인구수
가 많을수록 체감안전도가 낮다

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

Machine
Learning

체감안전도 예측

에이콘 아카데미 5조 발표