안전한 국민의 삶을 위한

시민이 공감하는 치안 체감안전도 예측 및 분석

Contents

C O N T E N T S 1

서 론

주제 선정 배경

기존 치안 체감안전도 측정

- 각 경찰서 관할 지역당 주민 일부에게 설문조사를 실시하여 치안 체감안전도를 측정함
- 1년에 상반기, 하반기로 나누어 총 2번 측정하며 점수는 6월과 12월에 발표하는 형식으로 진행하며, 2020년도부터는 연 1회 실시함
- 현장 치안인력 보강, 치안서비스 개선 등 다양한 경찰활동 및 치안정책을 개발하는데 활용하고 있음

주제 선정 배경

기존 치안 체감안전도의 한계점

- 2020년도 기준 1월부터 약 10개월 동안 일반국민 51,000명(관서별 200명)을 대상으로 무작위 전화 설문조사를 하는 방식으로 시간과 비용이 많이 소모됨
- 국민들의 즉각적인 치안 체감안전도를 알 수 없어 각 경찰서들의 피드백이 다소 늦게 이루어짐
- 경찰이 가지고 있는 자원과 지방자치단체나 주민들이 가진 자원들을 제대로 활용하지 못하여 데이터의 활용이 효과적으로 이루어지고 있지 않은 상황

따라서 치안만족도에 영향을 미치는 변수들을 찾고, 이를 활용한 예측 모델개발을 통해

체감안전도를 실시간으로 예측할 수 있는 체계를 마련하고자 함

C O N T E N T S 2

데 이 터

사용 데이터 설명

1. 기본 제공 데이터

: 데이터 분석을 위해 제공받은 데이터 전반 (2017 ~ 2020)

경찰청

2. 공공데이터포털 및 지역별 구청 포털

: 분석을 위한 추가적인 데이터 (2017 ~ 2020)

DATA 공공데이터포털 .GO.KR

3. KOSIS (Korean Statistical Information Service)

: 화재 발생 건수 데이터 (2017)

통계청

인구 데이터 전처리 – 관할서별 거주 인구

	sido	sgg_nm	age	date	popu_num	…	local_sx_rate
0	서울특별시	종로구	합계	2017	157277	…	96.3
1	서울특별시	종로구	0~4세	2017	4137	…	103.4
…	…	…	…	…	…	…	…
13822	서울특별시	중구	평균연령	2017	42.8	…	NaN
13823	서울특별시	중구	중위연령	2017	42.6	…	NaN

➔ 데이터 :

Sido, sgg_nm, age, date, popu_num, popu_male_num, popu_female_num, popu_sx_rate, local_num, local_male_num, local_female_num, local_sx_rate

sgg_nm : 종로구, 용산구, …, 마산회원구, 진해구, 중구 –〉 36개의 구로 이루어짐.
41개의 관할서별 담당 면적을 시각화 하여 대략적인 비로 나누어 관할서별 거주 인구수를 파악

분석에 필요한 sgg_nm, age, date, popu_num, popu_num, popu_male_num, popu_female_num을 뽑아 데이터 정제

age : 합계, 0~4세, 0세, …, 15세미만, 15~64세, 65세 이상, 평균연령, 중위연령
➔ 합계를 이용하여 전체 관할서별 거주 인구 파악
➔ 15세미만, 15~64세, 65세이상을 이용하여 취약계층인구과 경제활동인구 수를 파악

인구 데이터 전처리 – 관할서별 거주 인구

	year	name	popu_num	local_num	popu_male_num	popu_female_num	foreigner_num	vulner_popu_num	active_popu_num
0	2017	세종경찰서	276589.0	271299.0	139347.0	137242.0	5290.0	81129.0	195460.0
1	2017	진주경찰서	353209.0	347571.0	176672.0	176537.0	5638.0	98325.0	254884.0
...	...	...	...	...	...	...	...	...	...
121	2019	서울강북경찰서	303871.0	298525.0	147112.0	156759.0	5346.0	84901.0	218970.0
122	2019	서울노원경찰서	522480.0	516201.0	252662.0	269818.0	6279.0	136852.0	385628.0

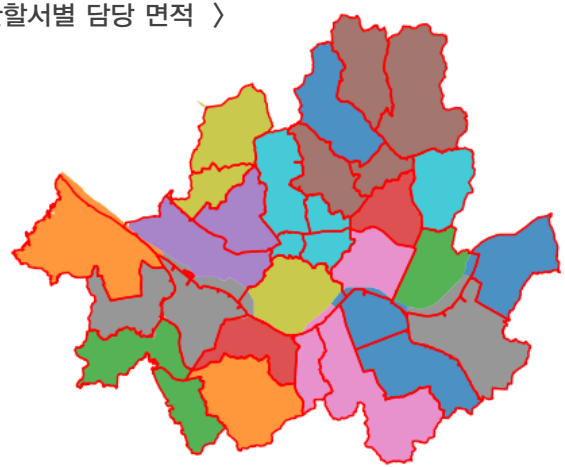
전체 인구수(popu_num) = age(합계) 총인구수
내국인 인구수(local_num) = age(합계) 내국인 인구수
남성 인구수(popu_male_num) = age(합계) 남성 인구수
여성 인구수(popu_female_num) = age(합계) 여성 인구수
외국인 인구수(foreigner_num) = age(합계) 총인구수 – age(합계) 내국인 인구수

취약계층인구(vulner_popu_num)
=15세미만 인구수 + 65세이상 인구수
경제활동인구(active_popu_num)
=15~64세 인구수

인구 데이터 전처리 – 관할서별 거주 인구(함수)

〈 실제 서울 지도 〉

〈 관할서별 담당 면적 〉



강남구 : 강남경찰서, 수서경찰서 담당

종로구 : 종로경찰서, 혜화경찰서 담당

중구 : 중부경찰서, 남대문경찰서 담당

성북구 : 종암경찰서, 성북경찰서 담당

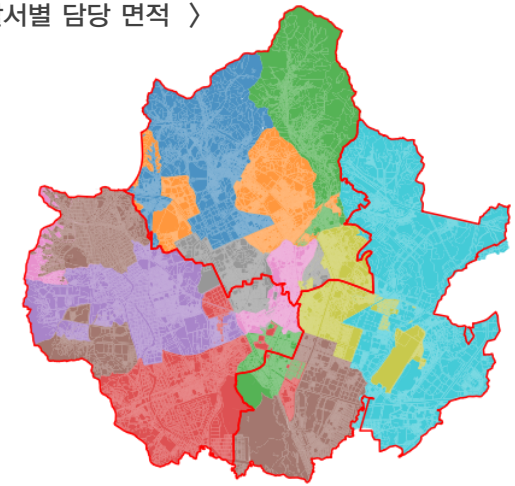
은평구 : 은평경찰서, 서부경찰서 담당

서초구 : 서초경찰서, 방배경찰서 담당

인구 데이터 전처리 – 관할서별 거주 인구(함수)

〈 실제 수원 지도 〉

〈 관할서별 담당 면적 〉



장안구 : 수원중부경찰서 담당

영통구 : 수원남부경찰서 담당

권선구 : 수원남부경찰서, 수원서부경찰서 담당

팔달구 : 수원중부경찰서, 수원남부경찰서,
수원서부경찰서 담당

인구 데이터 전처리 – 학력변수

	sido	sgg_nm	sx	age	inv_edu	num_in_2015	age_cat
24	서울특별시	종로구	남자	20–29세	계	487.0	20
25	서울특별시	종로구	남자	20–29세	초등학교	0.0	20
26	서울특별시	종로구	남자	20–29세	중학교	0.0	20
27	서울특별시	종로구	남자	20–29세	고등학교	66.0	20

	sgg_nm	inv_edu	num_in_2015
0	강남구	계	415123.0
1	강남구	고등학교	67927.0
2	강남구	대학교(2,3년제)	38989.0
3	강남구	대학교(4년제 이상)	209358.0
		…	
284	창원시 진해구	대학원(석박사 과정)	3764.0
285	창원시 진해구	받지 않았음(미취학 포함)	2585.0
286	창원시 진해구	중학교	10539.0
287	창원시 진해구	초등학교	8099.0

1. age_cat 변수를 생성해 연령대 카테고리를 생성하고 이를 기준으로 20대 이상 성인 데이터만 사용
2. 구(sgg_nm)와 학력변수(inv_edu)를 기준으로 groupby()함수를 활용해 학력에 따른 구별 인구수 집계
3. 타 인구 변수를 생성할 때 사용했던 기준과 동일한 기준을 적용해 관할서별 학력변수로 변환

☑ 19번 데이터는 5년에 한번 집계되는 데이터로 2015년 값만 주어졌으나, 학력은 경제적 수준, 직업유형 등과 함께 체감안전도에 영향을 미치는 중요한 변수로 알려져 있으므로 2015년 값을 매해 반복사용해 지역별 특성을 반영하는 변수로 활용하고자 함

인구 데이터 전처리 – 학력변수

	year	name	highedu	lowedu
0	2017	세종경찰서	0.45283	0.0945615
1	2017	진주경찰서	0.351681	0.109318
2	2017	창원서부경찰서	0.301793	0.0572083
3	2017	창원중부경찰서	0.362266	0.0425594
...				
160	2020	서울도봉경찰서	0.341758	0.0758287
161	2020	서울은평경찰서	0.364121	0.0738042
162	2020	서울강북경찰서	0.273791	0.102602
163	2020	서울노원경찰서	0.421741	0.0686943

학력 변수 산식)

Highedu : (경찰서별 대학교(4년제 이상) 인구 + 대학원(석박사 과정) 인구) / 경찰서별 조사자 합

Lowedu : (경찰서별 미취학 인구 + 초등학교 인구) / 경찰서별 조사자 합

- 지역별 특성이 두드러지도록
  고학력인구 비율에는 4년제 대학 이상의 학력 인구를
  저학력인구 비율에는 초등학교 이사의 학력 인구를 기준으로 사용

지리 데이터 전처리 – CCTV

	address	purpose	lat	lon	year	geometry	sp_purpose	police
0	세종특별자치시 조치원읍 교리	다목적	36.60	127.22962	2001	POINT(127.22962 36.60506)	안전	세종경찰서
1	세종특별자치시 조치원읍 남리	다목적	36.59	127.30245	2001	POINT(127.30245 36.59768)	안전	세종경찰서
2	세종특별자치시 조치원읍 남리	다목적	36.59	127.30049	2001	POINT(127.30049 36.559486)	안전	세종경찰서
3	세종특별자치시 조치원읍 명리	다목적	36.60	127.30130	2001	POINT(127.30130 36.60062)	안전	세종경찰서
4	세종특별자치시 조치원읍 상리	다목적	36.60	127.30314	2001	POINT(127.30314 36.60243)	안전	세종경찰서

〈 CCTV 목적 별 카테고리 〉

교통	안전	생활
차량방범	어린이보호	재난재해
교통정보수집	생활방범	쓰레기단속
교통단속	공원안전	기타
불법주정차	어린이안전	시설물관리
체납차량단속	방범	쓰레기무단투기
그린파킹	다목적	재난안전

〈 교 통 〉 〈 교통 외 〉

1. CCTV 데이터의 위도, 경도를 결합해 위치 정보를 담은 geometry 변수 생성
2. 관할서 경계 데이터의 MULTIPOLYGON 변수를 기반으로 각 보안등 위치가 어떤 관할서 영역에 포함되는지 라벨링한 police 변수 생성
3. CCTV의 목적에 따라 교통 / 안전 / 생활로 카테고리화해 사용 –〉 실제 데이터 분석에서는 교통 / 비교통으로 카테고리를 나눠 사용함

지리 데이터 전처리 – CCTV

〈 추 가 데 이 터 〉

	address	purpose	lat	lon	year	geometry	sp_purpose	police
0	서울특별시 강동구 길동	어린이보호	36.5408	127.147	2011	POINT(127.147 36.5408)	안전	서울강동경찰서
1	서울득별시 강동구 길동	어린이보호	36.5419	127.148	2014	POINT(127.148 36.5419)	안전	서울강동경찰서
...								
2806	Nan	어린이안전	37.6583	127.046	2005	POINT(127.046 37.6583)	안전	서울도봉경찰서
2807	Nan	어린이안전	37.6426	127.053	2005	POINT(127.053 37.6426)	안전	서울도봉경찰서

서울도봉경찰서 / 서울동대문경찰서 / 서울강동경찰서 데이터 누락 확인

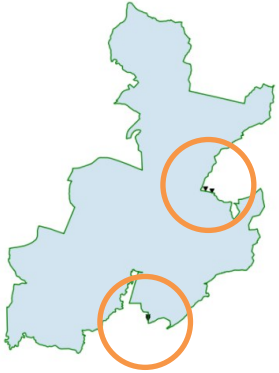
➔ 외부 데이터로 보완

외부 데이터도 기존과 같은 형태로 편집해 총 2807행 추가

관할서가 명확한 데이터를 사용했으므로 police 변수는 일괄 부여

지리 데이터 전처리 – CCTV

경찰서 값이 누락된 데이터들이 있어 확인한 결과 주소지 상 41개 경찰서 관할서에 포함되는 것이 맞으나
위도와 경도만을 기준으로 하는 geometry 변수 값의 차이로 포함되지 않았음
➔ 주소지 상 관할서와 시각화를 통해 확인한 관할서를 종합해 관할서 부여



	address	purpose		year	geometry	police	sp_purpose
35797	경기도 수원시 영통구	생활 방범	…	2012	POINT (127.06 37.26)	None	안전
35798	경기도 수원시 영통구	생활 방범		2012	POINT (127.06 37.26)	None	안전
38008	경기도 수원시 영통구	생활 방범		2019	POINT (127.04 37.23)	None	안전

…

예시) 주소지 상 각 포인트들은 수원중부경찰서에 포함되는 것이 맞으나
위치 관할서 경계에서 살짝 벗어나 있어 경찰서 값이 누락되는 현상 발생

CCTV 카테고리에 따라 총 3개의 변수 생성

	name	year	CCTV_safe
0	서울남대문경찰서	2017	209
1	서울혜화경찰서	2017	302
2	서울종로경찰서	2017	296
3	마산중부경찰서	2017	155
…			
160	서울송파경찰서	2020	2010
161	서울강북경찰서	2020	2192
162	서울광진경찰서	2020	3030
163	서울영드포경찰서	2020	3249

☑ 서울송파경찰서의 경우 2017~2019년은 집계되지 않고 2020년 값만 존재하는데, 이는 모델 학습에 저해요소로 작용할 것으로 판단해 2020년 값을 2017~2020년에 일괄 적용

범죄 데이터 전처리 – 사건 발생, 체포 건수

	name	crm	5m_crm_yn	crm_date
1	서울수서경찰서	위조외국통화행사	폭력	2017-01-01
2	서울영등포경찰서	사기	미분류	2017-01-01
3	서울서초경찰서	폭행	폭력	2017-01-01

기존 데이터는 각 경찰서가 담당한 범죄명과 5대 범죄에 들어가 있는지를 보여줌

그렇지만 범죄의 종류가 1000개가 넘고, 5대 범죄만으로 치안 체감만족도를 설명하기엔 무리가 있다고 판단함

	name	year	occur_theft	occur_traffic
1	마산동부경찰서	2017	796	1186
2	마산동부경찰서	2018	769	1012
3	마산동부경찰서	2019	693	1025

치안 체감안전도 설문조사에 포함된 범죄들과 그에 관련된 범죄 키워드 검색을 기준으로 하여 경찰서별 범죄 발생수를 카운트함

➔ 사건 발생과 체포 별 절도, 폭력, 강도, 살인, 교통 범죄, 경범죄, 집회 관련, 불법 광고물, 생활보장법 관련, 성범죄 변수 생성

범죄 데이터 전처리 – 시계열 예측의 사용

	name	year	occur_theft	occur_traffic
1	마산동부경찰서	2017	796	1186
2	마산동부경찰서	2018	769	1012
3	마산동부경찰서	2019	693	1025

	name	year	occur_theft	occur_traffic
1	마산동부경찰서	2017	796	1186
2	마산동부경찰서	2018	769	1012
3	마산동부경찰서	2019	693	1025
4	마산동부경찰서	2020	666	851

상당수의 데이터들은 이 2019년도 이전 값 밖에 존재하지 않은 것을 확인하였고, 2020년도 예측을 위해
데이터들을 사용하기 위해서 이전 년도의 데이터들로 차분을 이용한 시계열 예측을 실시함

범죄 데이터 전처리 – 시계열 예측의 사용

	year	occur_theft
1	2017	796
2	2018	769
3	2019	693

	year	occur_theft	d1c
1	2017	796	NA
2	2018	769	−27
3	2019	693	−76

	year	occur_theft	d1c	Forecast
1	2017	796	NA	NA
2	2018	769	−27	666
3	2019	693	−76	590

〈계산예시 : 범죄발생 마산동부경찰서 절도〉

2020년 절도 발생 예측값

차분은 시계열의 수준에서 나타나는 변화를 제거하여 시계열의 평균 변화를
일정하게 만드는데 도움을 주며 추세나 계절성이 제거되거나 감소됨
해당 시계열 예측을 범죄 발생, 체포건수, 자살건수, 각각의 인구 데이터에 적용함

분석 모델 – 모델 변수

5가지 안전도를 안정적으로 예측하기 위해 각 안전도 별로 예측 모델을 생성

〈반응 변수〉 – 치안 체감안전도

: 2020년 이전에는 1년에 상반기, 하반기로 나누어 총 2번 측정하였으나 2020년도부터는 연 1회 실시됨.
다른 설명변수들 관측 값을 상하반기로 나눌 경우 특징을 보여주지 못하는 경우가 많아 연 2회 실시된 설문조사의 체감안전도 값의 평균 사용

〈설명 변수〉

: 전처리를 통해 생성한 변수들 중 범죄 및 안전지표 변수들을 관련된 안전도 모델에 기본 변수로 배정

안전도 모델	기본 변수
절도폭력	절도, 폭력, 성범죄 발생/체포
강도살인	강도, 살인, 성범죄 발생/체포
교통	교통 사고, 교통 범죄 발생/체포
법질서	생활보장법 위반, 경범죄, 집회 관련, 불법 광고물 발생/체포, 자살
전반	교통 사고, 화재, 자살

공통 변수
CCTV, 단란주점, 담당 면적, 연도, 관할서번호, 보안등, 비상벨 (교통 모델 제외)

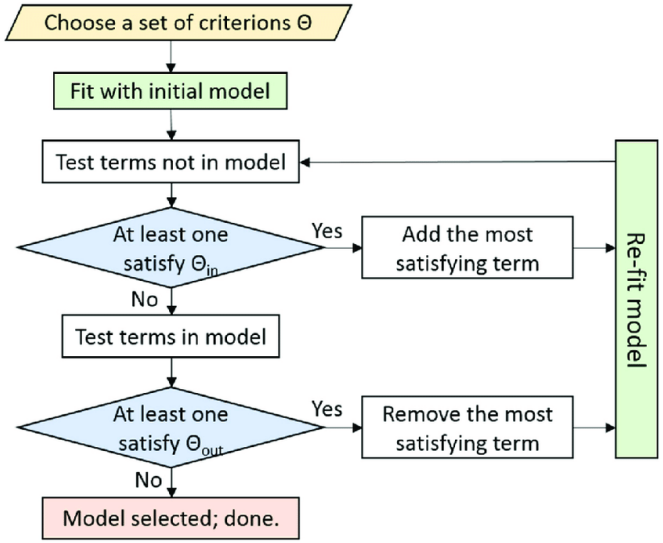
Stepwise 변수
인구 변수

* 전반 안전도 모델의 경우 전체 범죄 변수를 포함하고자 했으나 다중공선성이 클 것으로 우려돼 기본 변수로 배정하지 않고, stepwise selection을 통해 배정

C O N T E N T S 3

분 석 방 법 및 모 델

Stepwise



기존 회귀분석에서 모델은 학습이 지나치게 학습 데이터에 맞춰져
일반화 성능이 오히려 떨어지는 경우 존재

이를 해결하기 위한 방법 : 독립변수의 선택 ➔ stepwise

Step 1. 변수를 넣거나 제거할 때 boundary로 사용할 significance level을 정함
– Significance level :해당 값보다 높은 수치가 나온다면 해당 모델의 신뢰성이 낮아짐
Step 2. forward selection을 수행해서 선정된 변수 중 유의미한 변수만 남기고 제거
– 포함된 변수들을 학습시킨 후 나머지 변수들을 추가하며 다시 학습을 진행
Step 3. backward selection을 수행해서 선정된 변수 중 유의미한 변수만 남기고 제거
– 포함된 변수들을 학습시킨 후 p-value가 SL보다 높은 경우 해당 변수를 제거함
Step 4. 2, 3번 과정을 반복해서 수행 후 변수가 추가되거나 제거할 케이스가 없는 경우 종료

Stepwise

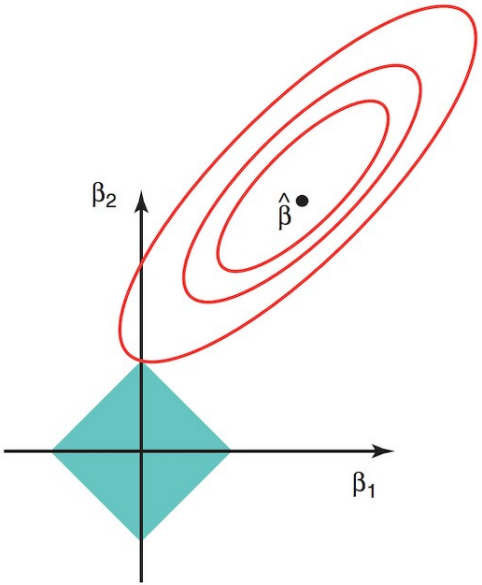
인구 관련 변수가 많기때문에 각 모델들마다 인구 변수를 다 넣는 것은 무리로 판단해 stepwise를 이용해 각 모델 중 가장 유의미한 변수들을 뽑고자 함

설문조사	영향 변수
절도폭력	저학력 인구 비율, 총인구 수, 취약계층 인구 수, 여성 인구 수
강도살인	저학력 인구 비율, 총인구 수, 취약계층 인구 수, 여성 인구 수
교통	영향 변수 없음
법질서	저학력 인구 비율, 경찰 수
전반	저학력 인구 비율, 여성 1인 가구 수

전반적 설문조사에서도 각종 사건들의 다중공산성이 심하다고 판단하여 stepwise를 통해 가장 유의미한 변수들을 뽑음

설문조사	영향 변수
전반	경범죄 발생, 교통 범죄 체포, 집회 범죄 체포, 교통 범죄 발생, 성범죄 체포, 살인 발생, 절도 체포

Lasso



일반 다중회귀식 : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

기존의 다중회귀분석은 주어진 샘플들의 특성들과 예측 값의 관계를
필요 이상으로 너무 자세하고 복잡하게 분석
➔ 이는 새로운 데이터가 주어졌을 때 예측력이 떨어지는 현상을 유발

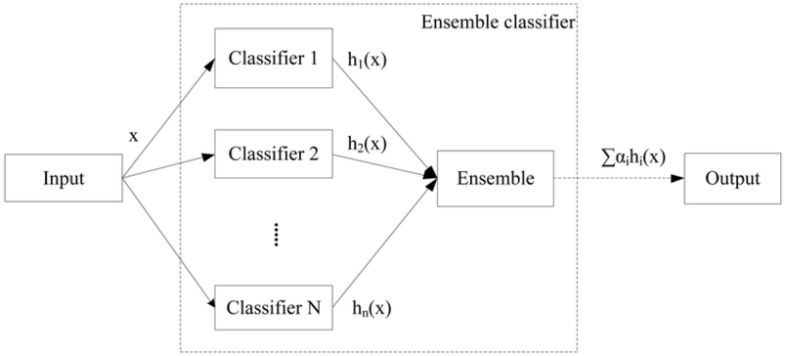
➔ 해당 현상을 예방하기 위해 LASSO 모델을 사용함

L1 Regularization

$$\text{Cost} = \sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{ij} W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

λ 값이 클수록 계수 추정치는 0에 가까워져 몇몇 유의미하지 않은 변수들에 더해
계수를 0에 가깝게 추정해주어 변수 선택 효과를 가져옴

Ensemble



〈ENSEMBLE〉

여러 독립적이고 다양한 결과들을 조합할 경우,
무작위 오차들이 서로 상쇄되어 좋은 결과를 가져올 수 있음

Ensemble은 한 종류의 학습 방법을 사용하지만,
학습 데이터를 조작하여 다양한 모델을 학습,
같은 알고리즘이지만 서로 다른 모수를 선택,
서로 다른 종류의 입력 변수를 사용

➔ Lasso에 Ensemble 기법을 적용시켜 정확도를 더 높이고자 함

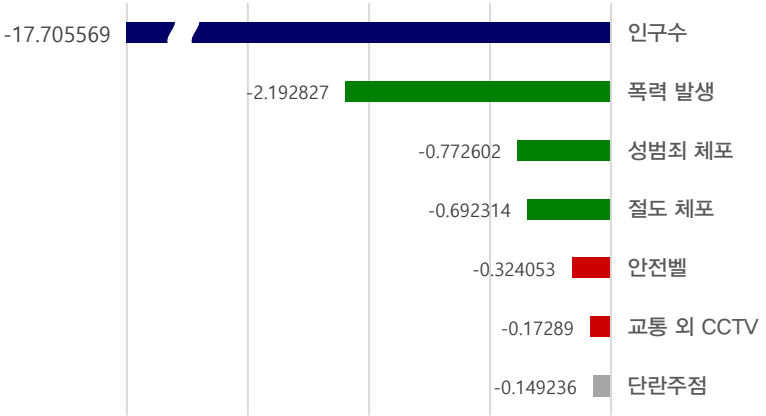
Train_set을 총 20개의 그룹으로 무작위로 분리하여
각 그룹의 최적인 회귀계수를 찾아내어 평균값을 구한 후 그 값을
최종 회귀분석 계수로 채택

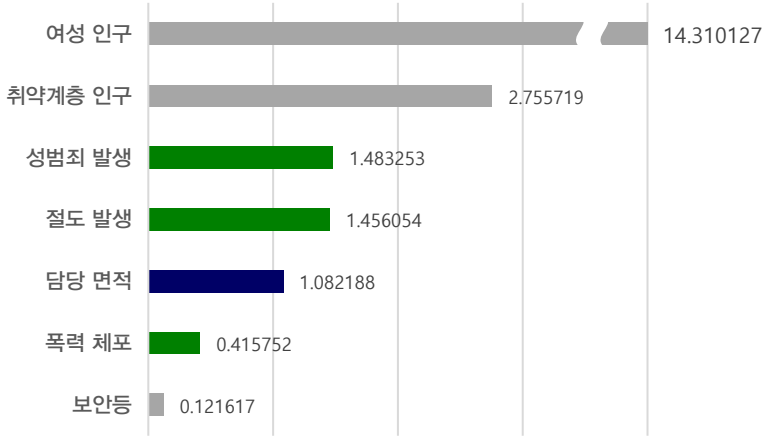
C O N T E N T S 4

결 론

분석 결과 (1) 절도폭력

“범죄 예방”

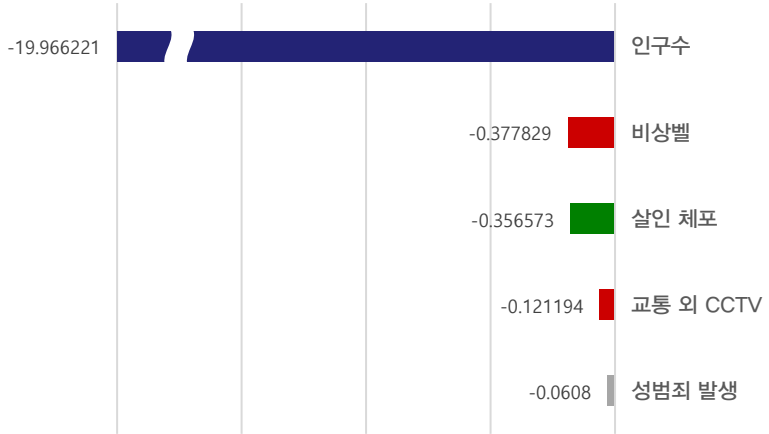




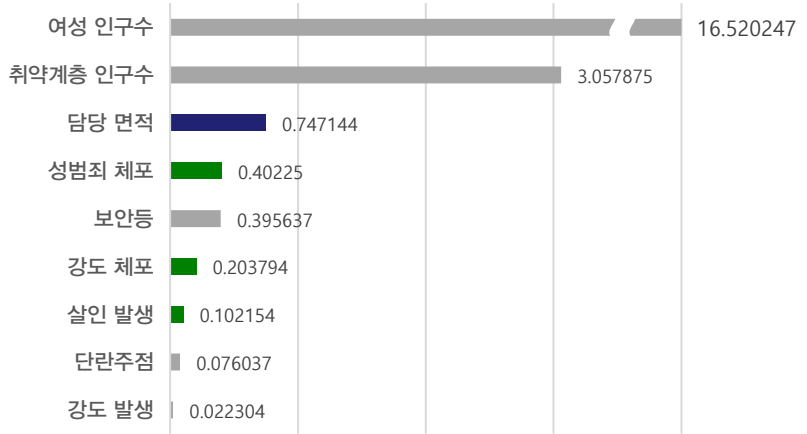
“불안감 해소”

“인력 배치”

분석 결과 (2) 강도살인



“검거 주력”

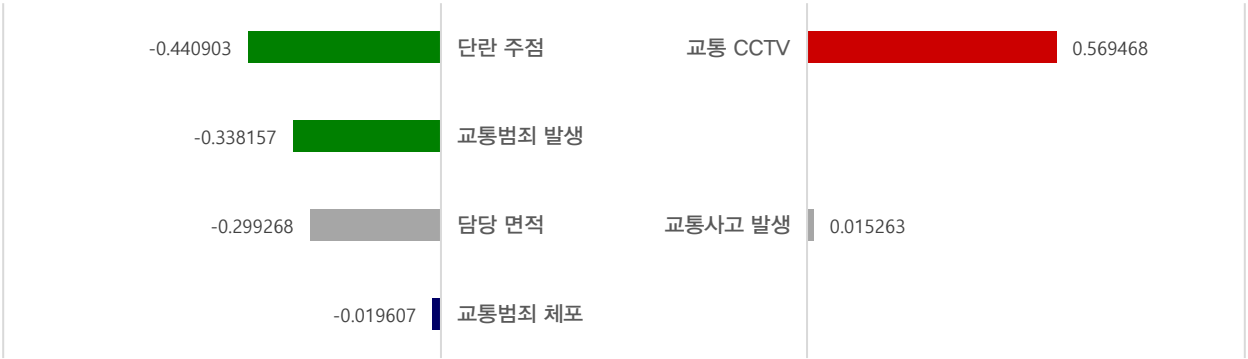


“불안감 해소”

“인력 배치”

분석 결과 (3) 교통

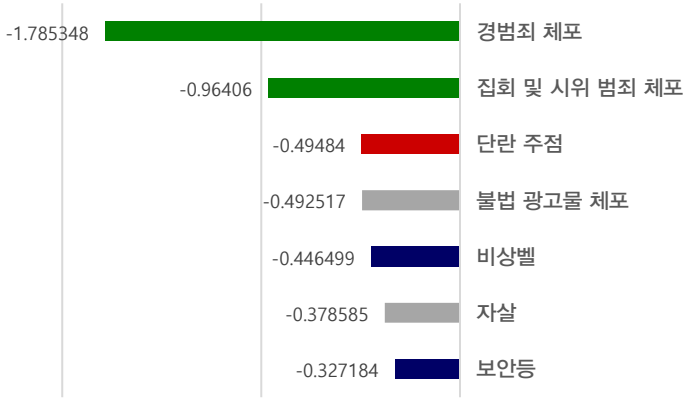
“음주운전”



“안전운전 유도”

“발생 원천 차단”

분석 결과 (4) 법질서



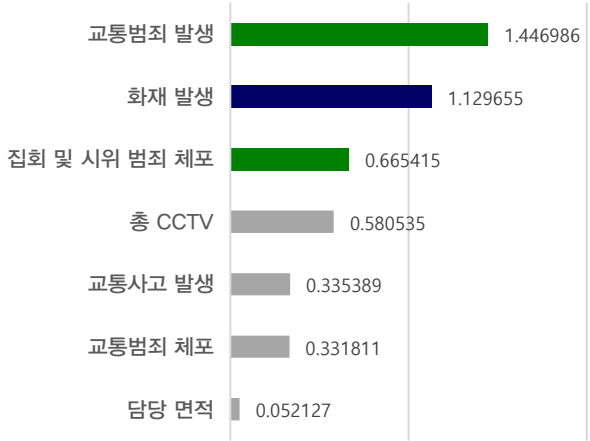
“현상 감소”

“불안감 해소”

“발생 억제”

분석 결과 (5) 전반적

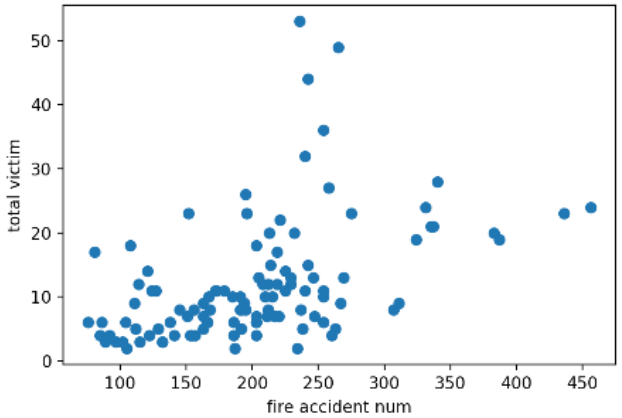
“범죄 발생”



“자살예방제도”

“후속 조치”

분석 결과 (5) 전반적



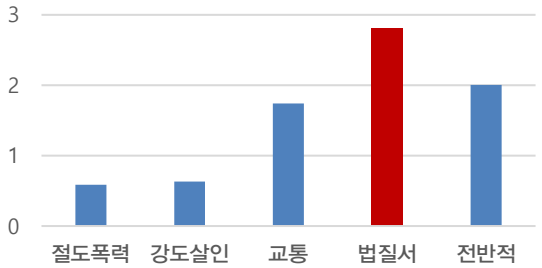
〈화재 변수〉

양의 회귀계수 값을 가져 예상과 다른 결과를 보임

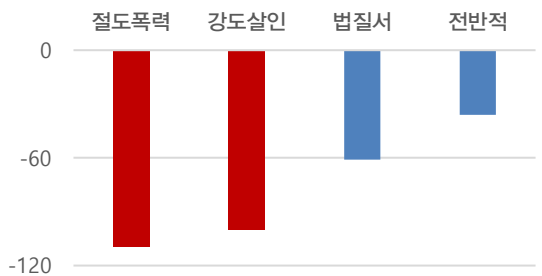
: 실제 현실에서 발생하는 화재의 경우 인명피해가 거의 발생하지 않는데, 이러한 특징에 따라 화재발생 자체에서 느끼는 불안감보다 화재 관련 후속조치를 통해 느끼는 안전도가 상대적으로 크게 작용한 것으로 예상

저학력 인구 변수, 연도 변수

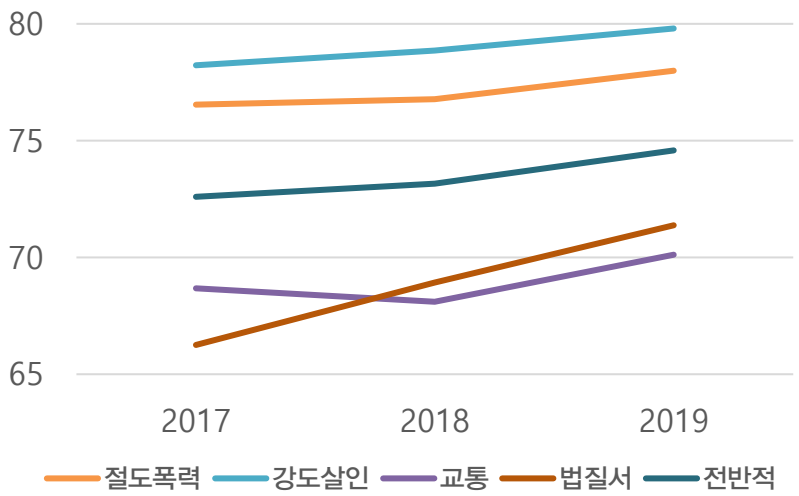
〈연도 변수의 모델별 회귀계수〉



〈저학력 인구 변수의 모델별 회귀계수〉



〈연도별 치안 체감안전도 증감추이〉



모델 평가

앞서 예측한 값들을 토대로 구한 MAE

1.3107

〈Public score 기준〉

기대효과

- 치안 체감안전도를 설문조사를 통하지 않더라도 데이터를 이용하여 구할 수 있으며 이에 따른 시간, 비용의 절감이 예상
- 각 경찰서 관할지역에서 발생하는 사건의 종류, 건수에 따라 실시간으로 체감안전도를 예측할 수 있어 각 경찰서별로 피드백을 보다 더 빠른 시일내에 마련할 수 있을 것으로 예상 (다양한 치안서비스 개선, 현장 치안인력 보강)
- 실시간으로 예측되는 체감안전도를 통해 국민들이 더 안전한 사회 속에서 살 수 있을 것으로 예상

한계점

- 관할서별 인구 수와 유동인구를 반영한다면 더 좋은 성능의 모델을 기대할 수 있을 것으로 예상
- 2020년부터 시작된 포스트 코로나 시대를 반영할 만한 변수가 존재하지 않음
- 각 모델마다 학력변수의 영향력이 꽤 큰 것으로 나타났지만 해당 조사는 5년마다 진행됐기 때문에 기대보다 살짝 더 높은 오차를 보여줌
- 112 신고 접수 데이터가 18년 5월부터 존재함. 치안 체감안전도에 중요한 영향을 끼칠 것으로 예상되나 데이터의 부족으로 활용하지 못함
- 2017년 교통사고 데이터의 부재로 교통 안전도 모델과 전반 안전도 모델에는 2018~2019년도 데이터만으로 학습이 진행되었으므로 데이터가 충분히 확보된다면 모델의 성능을 개선시킬 수 있을 것으로 예상
- 실질적으로 여성들에게 가장 큰 위협을 주는 성범죄의 항목이 아직까지 존재하지 않아 실질적인 여성들의 치안 체감안전도가 더 낮을 것으로 예상

Q & A