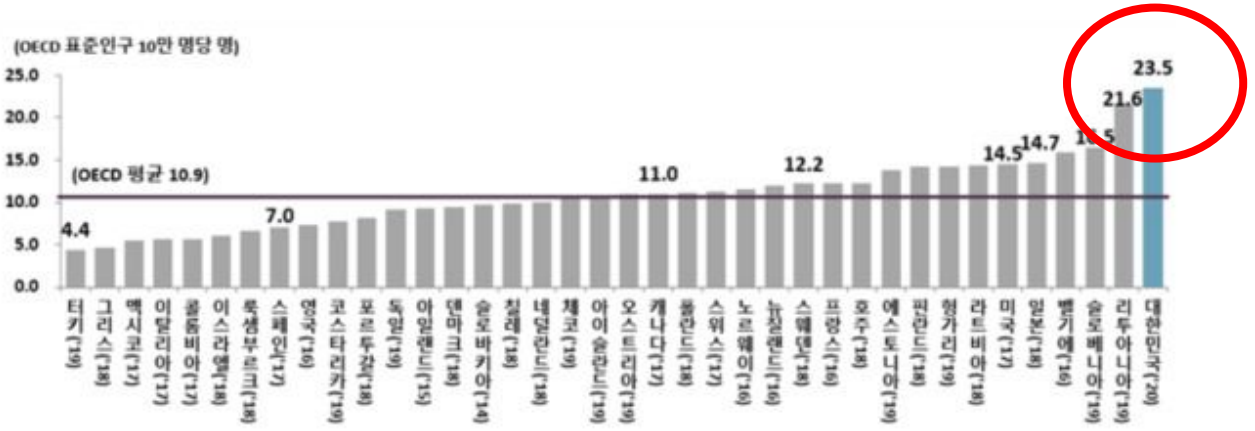
자살자 수 예측모델

AI_13 이창수

목차

1. 프로젝트 개요

OECD 국가 연령표준화 자살률 비교



• 자료: *OECD.STAT, Health Status Data(2021. 9. 추출), 우리나라 최근 자료는 OECD 표준인구로 계산한 수치임*
• OECD 평균은 자료 이용이 가능한 38개 국가의 가장 최근 자료를 이용하여 계산

1. 프로젝트 개요

분석의 목적

- 문제 상황 : 매년 1만2000~1만5000여명이 스스로 목숨을 끊고 있는 한국은 십수년 동안 'OECD 자살률 1위'라는 오명을 씻지 못하고 있음.
- 분석 목표 : Domain knowledge를 바탕으로 자살자 수를 예측하는 회귀모델을 만들고, 어떤 요인들이 자살자 수 증가에 영향을 미치는지 분석해 봄.

2. 데이터수집

자살의 원인

그동안 국내외 학계와 언론에서는 자살의 이유로 다양한 원인을 지목했음.

1. 경제적 이유
2. 기후적 이유
3. 질병적 이유
4. 계절적 이유

2. 데이터수집

- 지역별 자살자 수 데이터(통계청 mdis 마이크로데이터, 사망원인통계)
- 지역별 평균기온, 강수량, 습도, 일조량, 일사량(기상청 기후통계분석)
- 소비자물가지수(월), 가계부채(월), 실업자 수(월)(KOSIS국가통계포털)
- 코스피지수(일), 환율(일)(판다스 FDR라이브러리)
- 지역별 우울증 환자 수(건강보험심사평가원국민관심질병통계)

분석 기간 : 2011.1.1 ~ 2020.12.31

3. 전처리

1. pd.merge 기준 date+지역코드
2. 날씨데이터에 맞춰 자살데이터 지역코드 통합

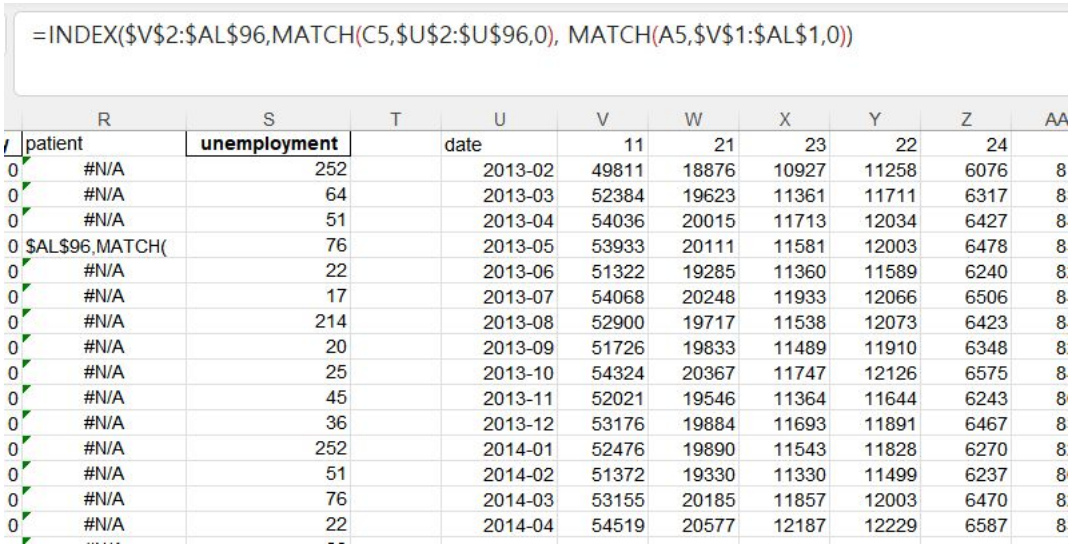
1.5.1 date dataframe 만들기

```
1  import datetime
2
3  index = pd.date_range(start = '2011-01-01', end = '2020-12-31', freq= 'D')
4
5  columns = ['weekday']
6
7  df_date = pd.DataFrame(index = index, columns = columns)
8  df_date = df_date.fillna(0)
9  df_date.index.name = 'date'
10 df_date.reset_index(inplace = True)
11 df_date.date = df_date.date.astype(str)
12
13 df_date.info()
```

executed in 41ms, finished 11:52:08 2022-05-22

3. 전처리

3. 복수 기준 데이터 : 엑셀 index + match + match 함수를 활용함.
월별/지역별 실업자 수, 월별/지역별 우울증 환자 수



=INDEX($V$2:$AL$96,MATCH(C5,$U$2:$U$96,0), MATCH(A5,$V$1:$AL$1,0))

	R	S	T	U	V	W	X	Y	Z	AA
/	patient	unemployment		date	11	21	23	22	24	
0	#N/A	252		2013-02	49811	18876	10927	11258	6076	8
0	#N/A	64		2013-03	52384	19623	11361	11711	6317	8
0	#N/A	51		2013-04	54036	20015	11713	12034	6427	8
0	$AL$96,MATCH(	76		2013-05	53933	20111	11581	12003	6478	8
0	#N/A	22		2013-06	51322	19285	11360	11589	6240	8
0	#N/A	17		2013-07	54068	20248	11933	12066	6506	8
0	#N/A	214		2013-08	52900	19717	11538	12073	6423	8
0	#N/A	20		2013-09	51726	19833	11489	11910	6348	8
0	#N/A	25		2013-10	54324	20367	11747	12126	6575	8
0	#N/A	45		2013-11	52021	19546	11364	11644	6243	8
0	#N/A	36		2013-12	53176	19884	11693	11891	6467	8
0	#N/A	252		2014-01	52476	19890	11543	11828	6270	8
0	#N/A	51		2014-02	51372	19330	11330	11499	6237	8
0	#N/A	76		2014-03	53155	20185	11857	12003	6470	8
0	#N/A	22		2014-04	54519	20577	12187	12229	6587	8

3. 전처리

4. 기존 데이터와 교차 검증

1.2.3 연단위 자살자 추이 + 데이터 전처리 검증

```
]:  1  df_tmp.groupby('year')['suicide'].sum()
```

executed in 14ms, finished 16:05:24 2022-05-23

```
t[24]: year
2011.0    15906.0
2012.0    14160.0
2013.0    14427.0
2014.0    13836.0
2015.0    13513.0
2016.0    13092.0
2017.0    12463.0
2018.0    13670.0
2019.0    13799.0
2020.0    13195.0
Name: suicide, dtype: float64
```

3. 전처리(data field)

target(1) : 'suicide' : 자살자 수
(지역별/날짜별)

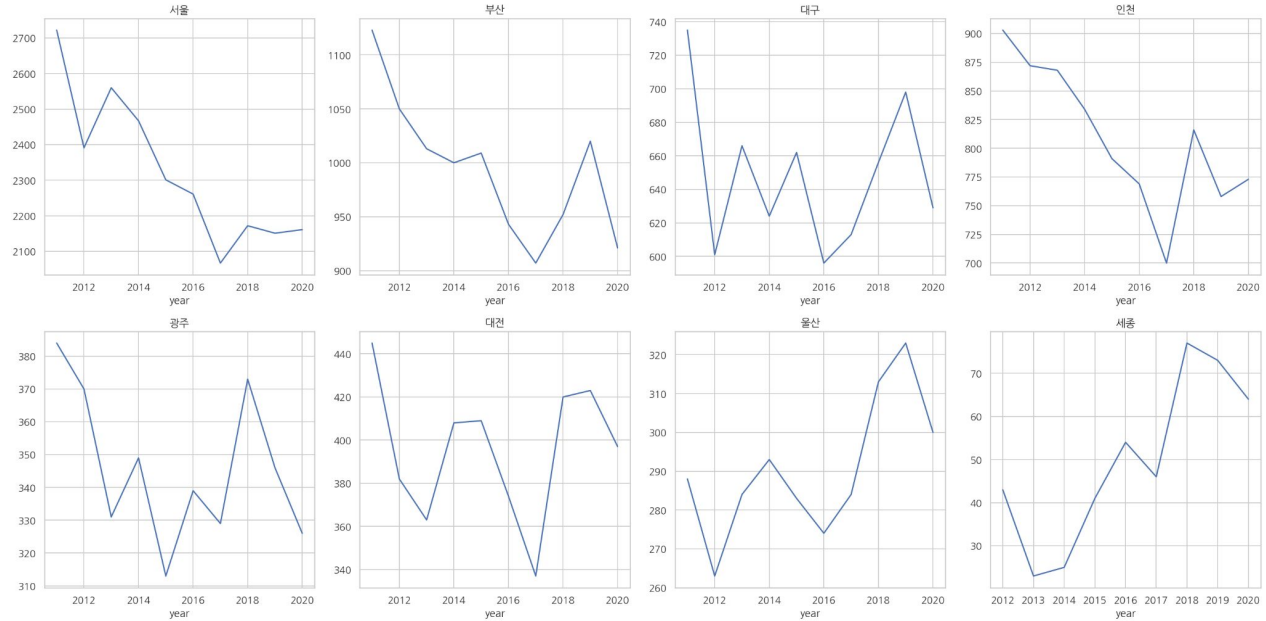
features(19) :

- 'loc_num' : 통계청 <사망원인통계> 지역 코드, 17개(A형, 시도 기준)
- 'date' : 자살 통계 집계일(지역 기준)
- 'year' : 연도
- 'month' : 월
- 'day' : 일
- 'weekday' : 요일
- 'loc_numm' : 기상청 <기후통계분석> 지역 코드, 9개
- 'loc_name' : 기상청 <기후통계분석> 지역명
- 'rain' : 일 평균 강우량(지역)
- 'temp' : 일 평균 기온(지역)
- 'hum' : 일 평균 습도(지역)
- 'sun' : 일 평균 일조량(지역)
- 'insola' : 일 평균 일사량(지역)
- 'exchange_rate' : 원/달러 환율(일)
- 'kospi' : 코스피지수 종가(일)
- 'price_index' : 소비자물가지수(월)
- 'house_debt' : 가계부채(월)
- 'patient' : 지역별 우울증 환자 수 (월, 통계청 지역코드 기준)
- 'unemployment' : 지역별 실업자 수 (월, 통계청 지역코드 기준)

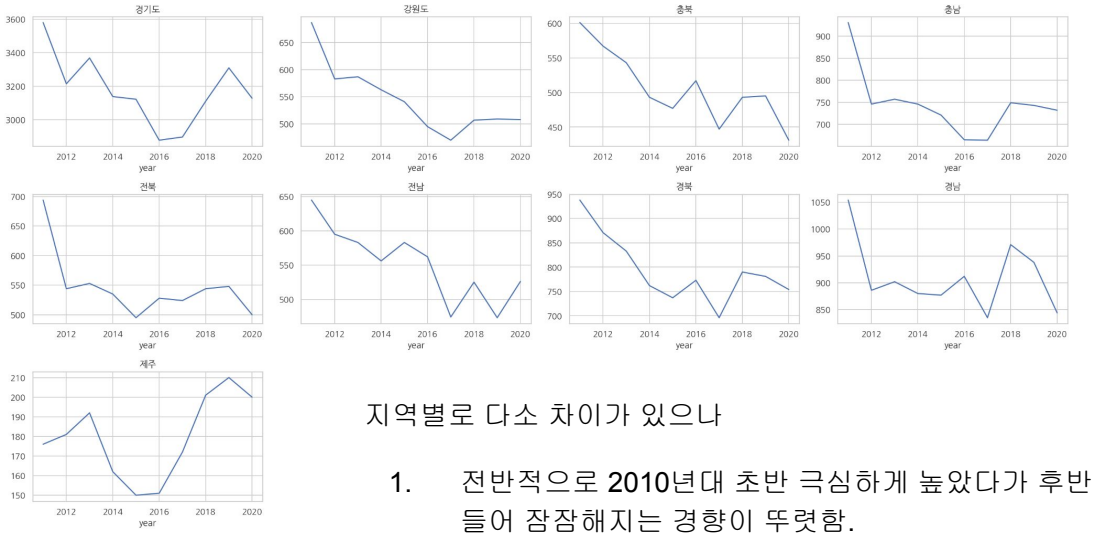
4. EDA

1. 월단위 자살자 추이
2. 연단위 자살자 추이
3. 요일별 자살자 수
4. 지역별 자살자 추이
5. 경제지표, 기후지표 시각화

4. EDA



4. EDA



지역별로 다소 차이가 있으나

1. 전반적으로 2010년대 초반 극심하게 높았다가 후반들어 잠잠해지는 경향이 뚜렷함.
2. 2015~2017년 즈음 최저치를 기록한 뒤 반등, 다시 하락하는 패턴 다수.
3. 제주, 울산, 세종 등은 최근 상승폭이 다른 지역과 비교해 뚜렷하게 크게 나타남.

5. 모델링

1. 기본모델로 돌린 뒤 모델 선택
2. LightGBM회귀모델 : tree기반 부스팅 알고리즘으로, 과적합에 다소 취약하지만 빠르고 성능이 좋음.
3. 기준모델 : 타깃 평균값(2.991)

```
1  dt_reg = DecisionTreeRegressor()
2  rf_reg = RandomForestRegressor()
3  xgb_reg = XGBRegressor()
4  lgb_reg = LGBMRegressor()
```

executed in 6ms, finished 21:46:10 2022-05-23

2 기준모델 ¶

```
1  baseline = [round(y_df.mean(),3)] * len(y_train)
2  print("기준 모델(평균) :",np.array(baseline))
```

executed in 25ms, finished 15:42:44 2022-05-24

기준 모델(평균) : [2.991 2.991 2.991 ... 2.991 2.991 2.991]

5. 모델링

4. 날짜/기후/경제/보건 feature들을 구분해 돌려봄

ver1 : features = ['loc_num', 'year', 'month', 'day', 'weekday', 'rain', 'temp', 'hum', 'sun', 'insola', 'exchange_rate', 'kospi', 'price_index', 'house_debt', 'patient', 'unemployment']

ver2 : features = ['loc_num', 'weekday', 'rain', 'temp', 'hum', 'sun', 'insola', 'exchange_rate','kospi', 'price_index', 'house_debt', 'patient', 'unemployment']

ver3 : features = ['loc_num', 'month', 'weekday', 'rain', 'temp', 'hum', 'sun', 'insola', 'exchange_rate','kospi', 'price_index', 'house_debt', 'patient', 'unemployment']

ver4 : 지역, 날짜, 금융, 실업자, 날씨, 우울증

모델의 r2_score, feature importance 등을 살펴봄.

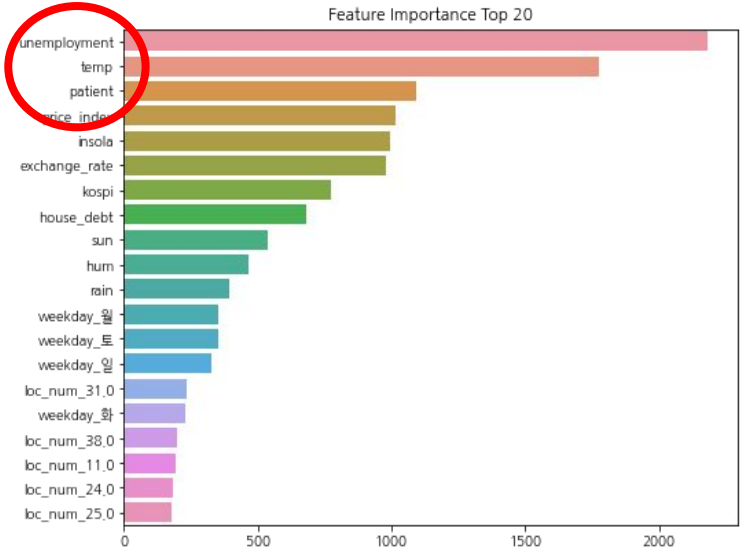
5. 모델링

5. 최종 모델링

- features : 요일을 제외한 날짜 데이터 삭제
- OneHotEcoding : 지역코드, 요일
- LGBMRegressor 모델
- RandomizedSearchCV를 활용해 하이퍼 파라메터 튜닝

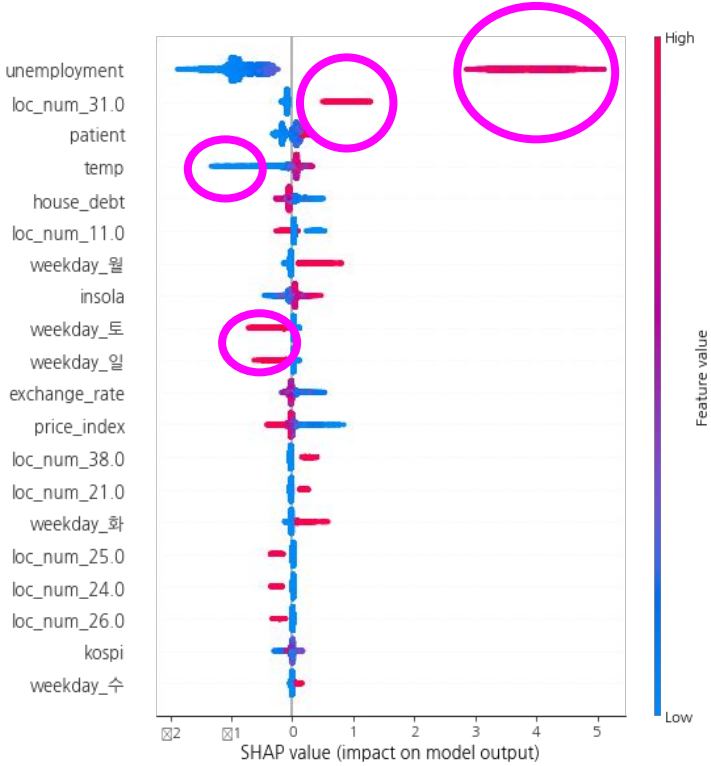
모델 성능
accuracy: 0.66729453815101
r2score: 0.6454682145730184
MAE: 1.1285417511852376
MSE: 2.456754444106057
MSLE: 0.12141958069598982
RMSLE 0.3484531255362618

6. 결과 분석

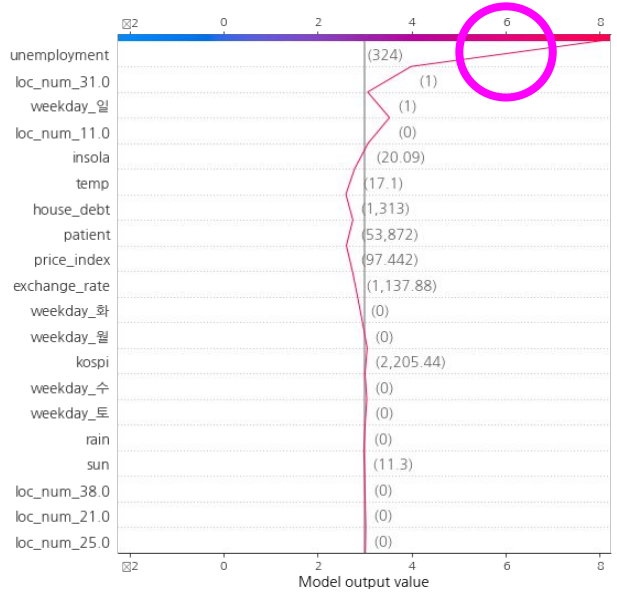


Weight	Feature
1.0270 ± 0.023	unemployment
0.0240 ± 0.002	loc_num_31.0
0.0200 ± 0.001	loc_num_11.0
0.0115 ± 0.001	temp
0.0071 ± 0.0013	house_debt
0.0062 ± 0.0014	weekday_토
0.0059 ± 0.0019	insola
0.0053 ± 0.0020	weekday_월
0.0051 ± 0.0020	price_index
0.0050 ± 0.0005	patient
0.0032 ± 0.0003	loc_num_38.0
0.0032 ± 0.0007	exchange_rate
0.0028 ± 0.0014	weekday_일
0.0025 ± 0.0005	weekday_화
0.0015 ± 0.0004	loc_num_25.0
0.0013 ± 0.0004	loc_num_26.0
0.0013 ± 0.0006	loc_num_24.0
0.0008 ± 0.0001	loc_num_21.0
0.0006 ± 0.0001	loc_num_34.0
0.0005 ± 0.0001	loc_num_37.0

6. 결과 분석

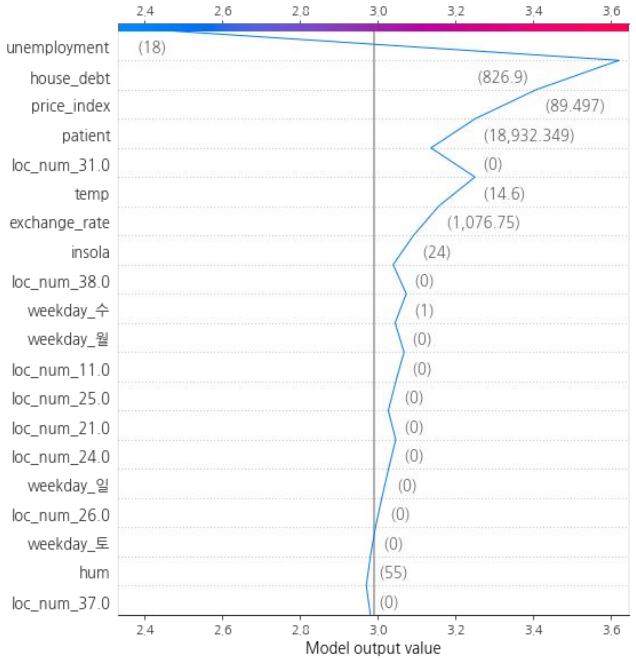


실제 :10명 예측 : 8.x명

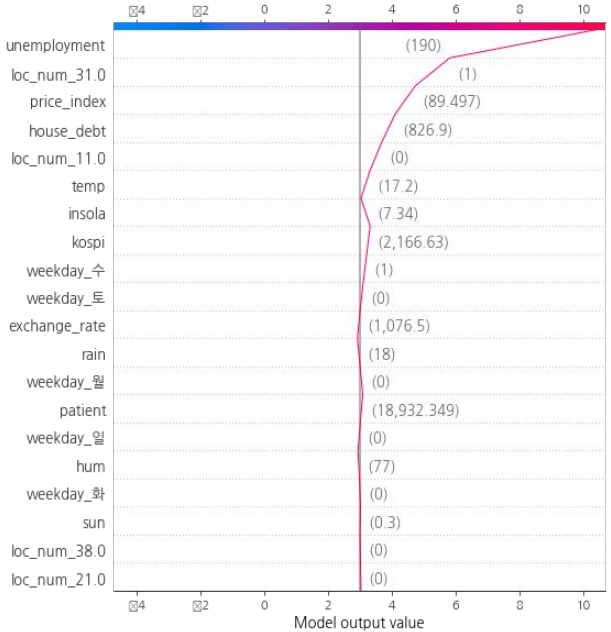


6. 결과 분석

실제 :3명 예측 : 2.x명



실제 :11명 예측 : 10.x명



6. 결론

- date 영향력 낮음
- 실업자 영향력 몹시 큼
- 경기/서울 여부
- 기온의 영향
- 주말 여부

- 자살이 계절적인 영향을 받을 것이라는 세간에 알려진 통념과 달리, date 변수들은 모델에 큰 영향을 끼치지 못했음.
- 자살은 지역 실업자 수에 가장 크게 영향을 받는 것으로 분석됨. SHAP value에 따르면 실업자 수가 많을수록 자살자 수가 커지고, 적을 수록 자살자 수가 적어짐. 적을 때보단 많을 때가 상대적으로 더 큰 영향을 미치는 것으로 보임.당초 인구 규모에 따라 자살자 수가 크게 영향을 받는 것으로 보았으나, 지역 변수(지난 10년간 지역별 인구가 매우 큰 폭으로 변동되지 않을 것이란 전제)보다 실업자 수에 더 민감하게 반응함.
- 지역적으로는 분석 대상 지역이 경기도와 서울인지 여부가 예측에 다소 영향을 미치는 것으로 파악됨. 경기도인 경우 자살자 수 상승 압력, 서울이 아닌 경우 자살자 수 하락 압력으로 작용.
- 날씨 feature에서는 기온의 경우, 기온이 낮아지면 자살자 수가 상대적으로 크게 줄어드는 경향 나타남. 일사량이 높을수록 상승, 낮을수록 하락.
- 요일별로는 토요일, 일요일인 경우 자살자 수 하락 압력, 월요일, 화요일인 경우 자살자 수 상승 압력으로 작용.
- 물가지수, 가계부채는 음의 상관관계, 환율은 양의 상관관계, kospi는 거의 영향을 미치지 못한 것으로 나타남. 추가적인 분석이 있어야 할 것으로 보임.