미세먼지 유발 영향인자 분석 및 발생량 예측

A반 2조 강지훈

과제 개요

1. 과제 정의: 기상 데이터를 통한 미세먼지 예측

- 주요 내용, 가설 목록, 목표 변수 설명

2. 데이터 수집 및 전처리

- 목표변수의 분포
- Feature Engineering
- 이상치 / 결측치 처리

3. EDA로 유의인자 찾기

- YearMonth / Season
- Rain / Humidity
- Wind / Wind Dir
- NO2, SO2, CO, O3

4. Multiple Linear Regression

- 모델 구현, 회귀 계수, 성능 평가

5. Decision Tree Regressor

- 모델 구현, 변수 중요도, 성능 평가

6. Ensemble

- Random Forest
- Gradient Boosting

7. Vital Few 도출 및 결론

- 유의 인자 도출, 모델 종합 평가

8. 피드백 / 개선 방향

1. 과제 정의: 기상 데이터를 통한 미세먼지 예측

미세먼지(PM10)란?

미세먼지를 관측하는 단위인 **PM10은 "직경 10마이크로미터 이하의 먼지"를 뜻한다.**

- 미세먼지는 단일 물질이 아니라 다양한 생성과정에 의해 만들어지는 미세한 고체와 액체 입자로, 대기 중에 떠 있는 아주 작은 먼지다.
- 생성원에 따라 1차/2차 미세먼지로 나뉜다.
    - 1차: 연료연소시설의 굴뚝, 자동차 배기구 등
    - 2차: 대기 중에서 황산화물, 질산화물, 암모니아가 화학 반응을 통하여 생성하는 유기물질

미세먼지 예측에서 기상정보의 중요성

미세먼지를 잘 예측하려면 미세먼지 관측 데이터가 많아야 하고, 미세먼지 관측 데이터가 많으려면 미세먼지 측정소가 많아야 한다.

- 그러나.. 실상은 미세먼지 측정소는 특정 지역에 집중되어 있고; 측정소가 희소한 지역에서는 미세먼지 농도 예측이 어렵다 => 이로 인해 지역 피해가 발생한다 (임준묵, 2019).
- 이에 반해 기상청의 기상관측장비는 전국 곳곳에 위치해 있고, 오랜 기간 동안 실시간으로 쌓인 데이터가 많다.
    - **만약 기상정보로 미세먼지를 예측할 수 있다면 관측재원(측정소)이 없는 곳에서도 미세먼지 농도 예측 / 피해 예방이 가능할 것이다.**

가설 목록

미세먼지에 대해 떠도는 속설 + 전문가들이 밝힌 연구내용들은 다음과 같은 것들이 있다.

- **비가 많이 오면 먼지가 씻겨 내려가서 미세먼지 농도가 낮아진다**. => 강수가 많은 여름철(7-9월)이 강수가 적은 겨울철/봄철(12-5월)보다 미세먼지 농도가 높을 것이다.

- **중국의 대기오염 때문에 중국에서 바람이 불어오면 국내 미세먼지 농도가 높아진다** => 즉, 풍향이 서쪽일 때가 동쪽일 때보다 미세먼지 농도가 높다.

- **풍속이 빠르면 먼지들이 빠르게 쓸려가서 미세먼지 농도가 낮아진다.** => 풍속이 낮을 경우 대기정체로 인해 미세먼지 농도가 높아질 것이다.

- **2차 미세먼지 생성원인 황산화물, 질산화물 등의 대기 중 농도가 높을수록 미세먼지 농도도 높을 것이다** => NO2와 SO2가 높을수록 미세먼지 농도가 높을 것이다.

데이터 수집 방법

- 주어진 데이터의 양이 너무 부족하다고 생각해서(대부분의 모델이 전혀 유의미한 설명력을 갖지 못할 정도로), 추가적인 데이터를 탐색했다.

- 주어진 데이터가 '서울' 지역이라는 건 알고 있었지만, '에어코리아'에서 제공하는 데이터는 서울 전체 평균이 아니라 지역구별로 데이터가 나누어져 있었다. 지역구별 데이터와 기존 갖고 있던 데이터와 비교해가며 가지고 있던 데이터가 서울 '중구' 데이터라는 것을 알게 됐다.

- 2019년7월~12월 데이터의 경우 에어코리아에서 제공하는 서울 중구 데이터와 기존 데이터와의 미세먼지 농도(PM10) MSE를 계산해보니 약 0.5 이하였다.

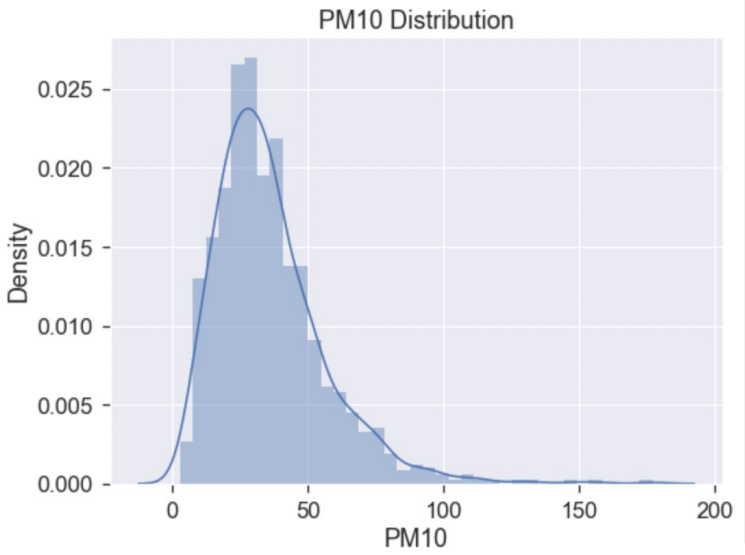
그래서 2018~2020년 전체 데이터를 쓰기로 했다.

기상자료의 경우 '기상자료개방포털'에서 서울 중구의 2018~2020년 기상데이터를 다운받았는데, 변수가 상당히 다양했다 (59개). 어떤 변수를 데이터셋에 합칠지 고민하다가, 임준묵(2019)의 연구를 참고해서 기존에 있던 기상변수인 **'기온, 강수량, 풍속, 풍향, 습도, 현지기압, 적설, 전운량' 등 8개에 더해 '평균지면온도, 강수계속시간, 평균이슬점온도, 안개계속시간'이라는 4개 변수를 추가하기로 했다.**

2. 데이터 수집 및 전처리

목표 변수 탐색

- PM10의 분포는 평균 36, 표준편차 20에 약간 right-skewed된 분포를 보여준다.
- 대부분의 값이 20~50 사이에 위치한다.
- 미세먼지 기준을 좋음(0~30), 보통(31~80), 나쁨(81~)로 나누면, 좋음이 474건, 보통이 525건, 나쁨이 34건이다.



설명변수 목록

- MeasDate: 측정일시
- NO2: 이산화질소
- O3: 오존
- CO: 일산화탄소
- SO2: 아황산가스
- TEMP: 온도
- RAIN: 강수량
- WIND: 풍속
- WIND_DIR: 풍향
- HUMIDITY: 습도
- ATM_PRESS: 기압
- SNOW: 적설량
- CLOUD: 전운량
- 지면온도
- 강수계속시간
- 평균이슬점온도
- 안개계속시간

결측치

- RAIN, SNOW, 강수계속시간, 안개계속시간 null값은 0
- NO2, O3, SO2, PM10, ATM_PRESS는 결측치가 적으므로 fillna(pad)적용
- CO는 PM10과 높은 상관관계를 지닌 유의 변수인데, 결측치가 꽤 많았다. (57개)
  - 그래서 NO2,O3,SO2,ATM_PRESS를 이용해서 CO를 예측하는 선형회귀모형을 만들어서 예측값으로 결측치를 채웠다.

MeasDate	0
NO2	3
O3	3
CO	57
SO2	3
PM10	1
TEMP	0
RAIN	426
WIND	0
WIND_DIR	0
HUMIDITY	0
ATM_PRESS	1
SNOW	633
CLOUD	0
지면온도	0
강수계속시간	641
평균이슬점온도	0
안개계속시간	1021

2. 데이터 수집 및 전처리

EDA를 통해 도출한 Vital Few

- **CO 결측치를 채우기 위한 선형회귀모델:** CO와 높은 상관을 가지는 NO2, O3, SO2, ATM_PRESS를 이용해 CO를 예측하는 선형회귀모형을 만들었다. 그리고 예측값을 이용해 결측치를 채웠다.

- **계절별 혹은 주기별 미세먼지 추이를 탐색하기 위해 연월(YearMonth) 변수**를 추가했다.

- **1~360 범위의 숫자로 표기된 '풍향'을** 북동, 북서, 남동, 남서, 총 4가지로 범주화했다.

- **공식적인 미세먼지 분류 기준에 따라** PM10이 30 이하면 '좋음', 80 이하면 '보통', 80 이상이면 '나쁨'으로 범주화했다.

- **봄, 여름, 가을, 겨울** 총 4가지의 계절 변수를 추가했다.

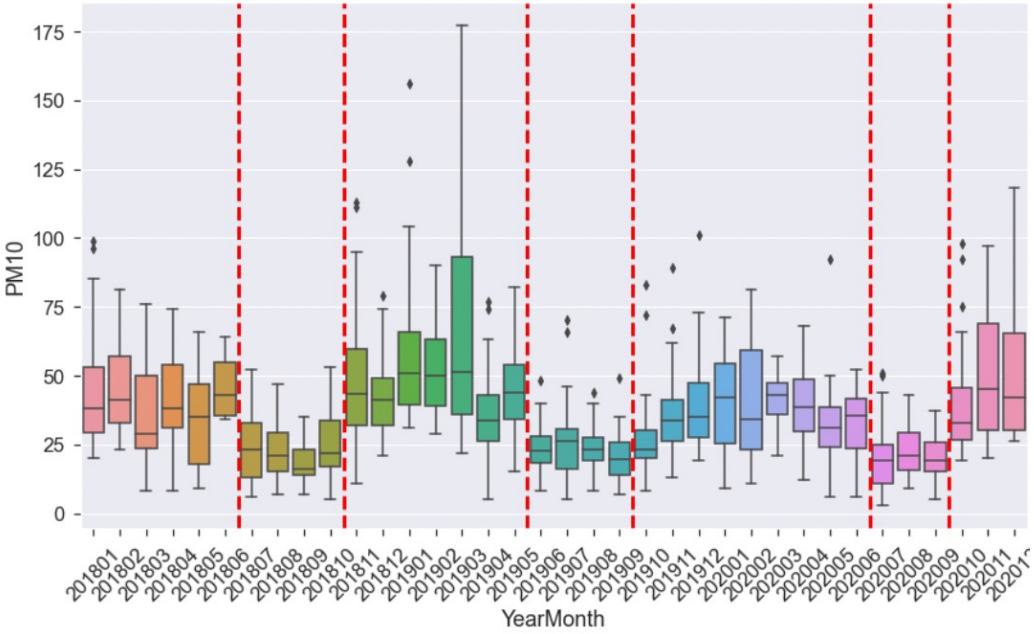
최종 변수

- **MeasDate: 일시**
- **YearMonth: 연월 (MeasDate의 파생)**
- **SEASON: 계절 (MeasDate의 파생)**
- **PM10: 미세먼지 농도 (ppm)**
- **PM_CAT: 미세먼지 상태 기준 (좋음0, 보통1, 나쁨2)**
- **NO2: 이산화질소 (ppm)**
- **O3: 오존 (ppm)**
- **SO2: 아황산가스 (ppm)**
- **CO: 일산화탄소 (ppm)**
- **TEMP: 평균기온 (섭씨)**
- **지면온도: 지면 평균온도 (섭씨)**
- **평균이슬점온도: 평균 이슬점온도 (섭씨)**
- **ATM_PRESS: 현기지압 (hPa)**
- **WIND: 풍속 (m/s)**
- **WIND_DIR: 풍향 (16방위를 4범주화시킴)**
- **RAIN: 일강수량 (mm)**
- **강수계속시간: (hour)**
- **HUMIDITY: 습도 (%)**
- **SNOW: 적설량 (mm)**
- **CLOUD: 전운량**
- **안개계속시간: (hour)**

3. EDA로 유의인자 찾기: YearMonth / Season

계절별 미세먼지 농도 변화

- 201801~202012까지 총 3년의 데이터를 시각화해보니, 5월~9월 사이에 미세먼지 농도가 눈에 띄게 낮은 것을 볼 수 있다
  - 정말 계절별로 PM10 농도가 유의미하게 다른지 확인하기 위해 ANOVA 분석을 했다.



ANOVA 분석

- "봄, 여름, 가을, 겨울" 간 미세먼지 농도 평균이 다른지 확인해보니 => 분석 결과 F 통계량 62.264, p-value는 0에 수렴한다.
  - 네 집단 간의 PM10 평균 사이에 통계적으로 유의미한 차이가 있다

Tukey's HSD 사후검정

- REJECT가 모두 True이므로, 서로 평균이 같은 그룹이 전혀 없는 것이다. 계절을 기준으로 집단을 묶으면, PM10이 모두 유의미하게 달라진다.
  - **분명히 계절이 PM10에 영향을 주는 것 같다.**

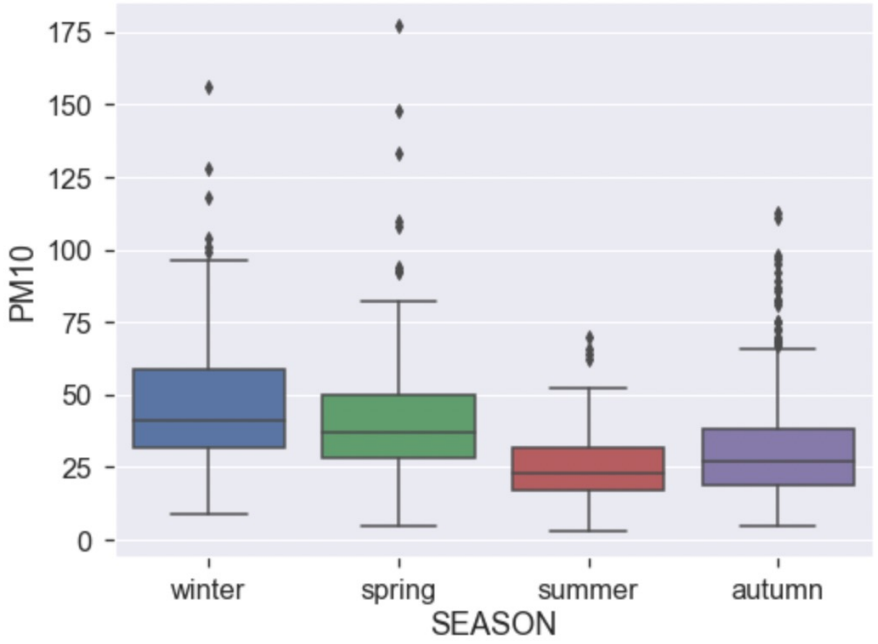
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
autumn	spring	9.0676	0.001	4.7452	13.3899	True
autumn	summer	-6.5515	0.001	-10.8184	-2.2846	True
autumn	winter	14.3237	0.001	10.1357	18.5117	True
spring	summer	-15.6191	0.001	-20.029	-11.2092	True
spring	winter	5.2561	0.01	0.9225	9.5897	True
summer	winter	20.8752	0.001	16.5969	25.1535	True

3. EDA로 유의인자 찾기: Season / Rain

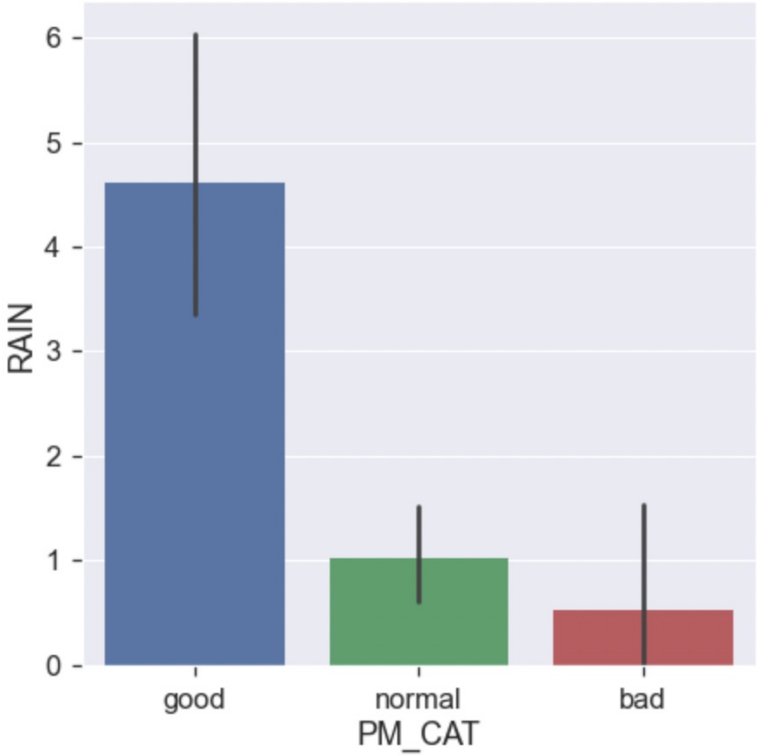
계절별 미세먼지 농도 변화 - Boxplot

- 서로 비슷해보이지만, Tukey's test 결과에 따라 이 네 집단 사이에 모두 통계적으로 유의미한 차이가 있다고 주장할 수 있다.
  - **겨울이 제일 미세먼지 농도가 높고, 여름이 제일 낮다.**



RAIN

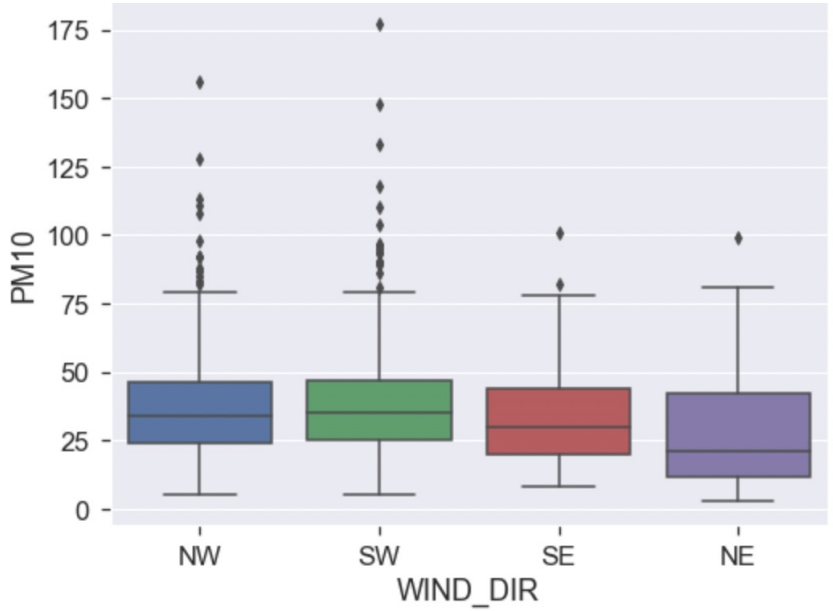
- 미세먼지 농도와 강수량의 관계를 시각화해보니, 미세먼지 농도가 '좋음'일 때는 '보통'이나 '나쁨'일 때보다 강수량이 약 3~4배 높았다.

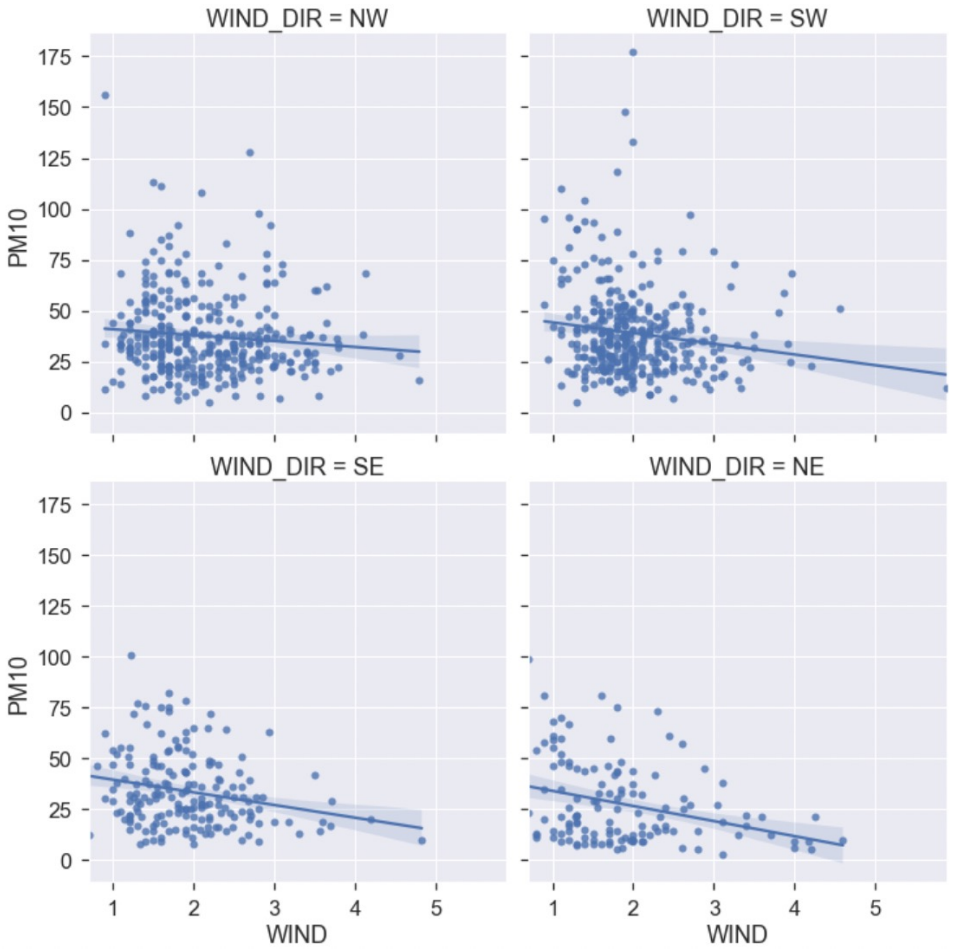


3. EDA로 유의인자 찾기: Wind, Wind_dir

풍향에 따른 미세먼지 농도 변화

- 풍속+풍향을 4개 plot으로 나눠서 PM10과의 관계를 봤는데,
  - 4개 plot 모두 다 풍속이 강해질수록 PM10가 미세하게 낮아지는 경향을 보였다. => **풍속이 강할수록 대기정체가 일어나지 않아서 미세먼지 농도가 낮아진다는 가설과 일치한다.**
- 풍향이 서쪽이고 풍속이 약할 때는 미세먼지 농도가 더 높길 기대했는데, 그래프만으로는 동풍과 큰 차이를 발견하기 힘들었다.



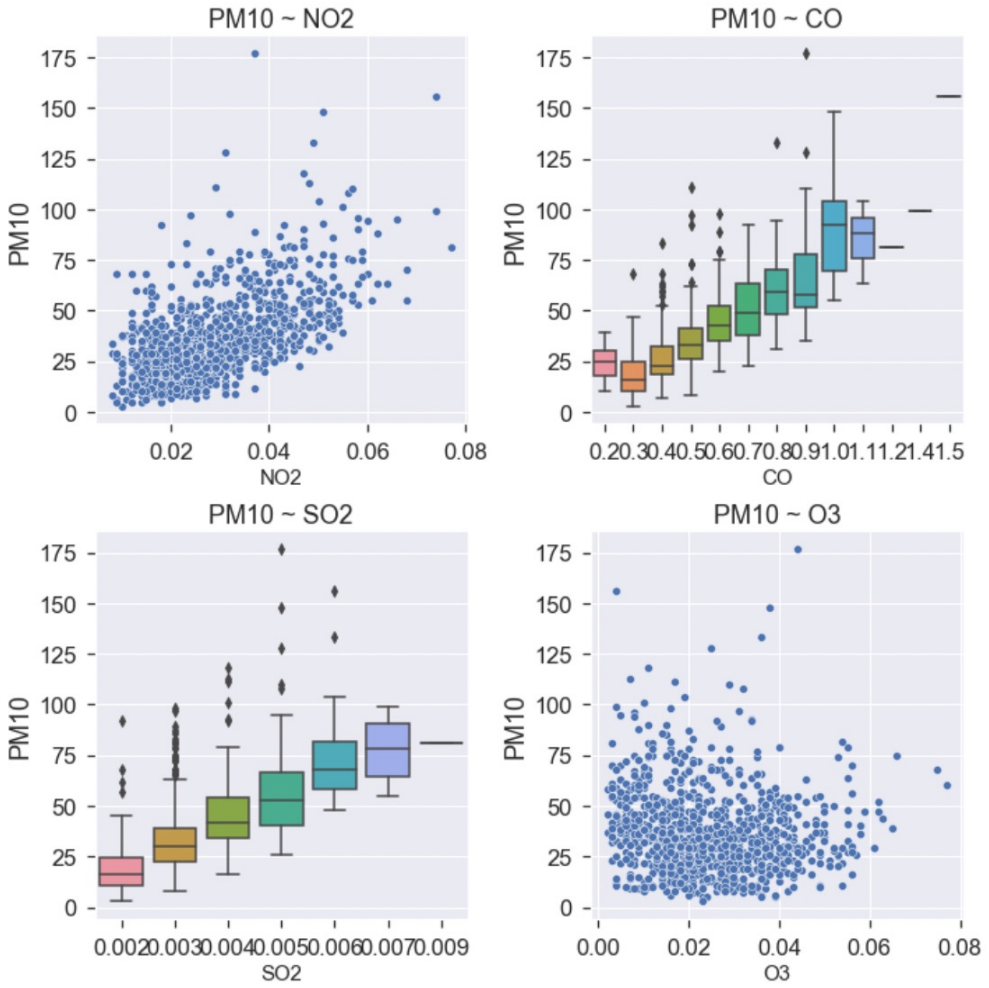


3. EDA로 유의인자 찾기: NO2, SO2, CO, O3

NO2, SO2, CO, O3와 미세먼지의 관계

- PM10과 NO2, SO2, CO는 모두 강한 양의 상관관계를 보였다.
- 다만 오존(O3)는 PM10과 큰 상관관계르 보이진 않았고, 오히려 O3는 NO2와 CO와 음의 상관관계를 보였다.

	PM10	NO2	SO2	CO	O3
PM10	1.000000	0.593982	0.604055	0.726600	-0.106240
NO2	0.593982	1.000000	0.645284	0.854938	-0.437544
SO2	0.604055	0.645284	1.000000	0.632421	-0.104308
CO	0.726600	0.854938	0.632421	1.000000	-0.430575
O3	-0.106240	-0.437544	-0.104308	-0.430575	1.000000



3. EDA로 유의인자 찾기: EDA 결론

EDA를 통해 도출한 Vital Few

- 겨울에 미세먼지가 제일 높고, 여름에 미세먼지가 제일 낮다(Time Series). 계절간 미세먼지의 농도 차이는 통계적으로 유의하다(ANOVA+Tukey's test).

- 미세먼지 농도가 Good이면 normal이나 good일 때에 비해 3배 이상 강수량이 높다 (box plot).
    - 하지만 전체 데이터를 기준으로 PM10과 일강수량의 상관관계는 낮다.

- PM10과 NO2, SO2, CO는 모두 강한 양의 상관관계(0.6이상)를 보였다 (correlation table).
    - 오존(O3)은 PM10과는 큰 상관관계가 보이지 않았고, 오히려 O3는 NO2와 CO와 음의 상관관계(-0.4)가 나타났다.

EDA를 통해 도출한 Trivial Many

- 풍속이 강할수록 PM10가 미세하게 낮아지는 경향이 있다 (lmplot).
    - 풍향이 서쪽일 때 미세먼지가 높길 기대했지만, 그래프만으로는 식별이 어려웠다.
    - NW-SW 그리고 NW-SE 간에는 PM10의 유의미한 차이가 없었다 (ANOVA+Tukey's test.)
    - 서풍과 동풍 간에 유의미한 차이가 없다는 통계결과는 최초 가설과 상반된다.

4. Multiple Linear Regression

모델링

- 모든 변수를 이용해서 다중회귀모델 만들었을 때 수정 R^2이 0.648이 나온다.
- RAIN, SNOW, CLOUD, 안개계속시간, NO2, (SEASON-Winter, WIND_DIR-SE)이 p-val 0.05 이상이다.
- '겨울'인 경우 계절이 PM10에 영향을 미치지 않는다 & 풍향이 남동풍(SE)일 경우 PM10에 영향을 미치지 않는다.

변수 제거 (p-value, VIF)

- 오른쪽의 p-value가 0.05 이상인 변수를 제거한 후, VIF가 10 이상인 변수도 추가로 제거했다. => 그리고 다시 다중회귀모형을 구현해서 p-value가 0.05 이상인 변수를 제거했다.
- 최종적으로 남은 가장 유의한 변수들은 다음과 같다:
  - **계절, 풍향, 오존, 일산화탄소, 풍속, 강수계속시간**
- 최종 6개 변수로 모델링한 결과 수정 R^2이 0.632가 나왔다.

Intercept	-361.9505	106.352	-3.403	0.001	-570.646	-153.255
C(SEASON)[T.spring]	6.7655	1.372	4.933	0.000	4.074	9.457
C(SEASON)[T.summer]	3.9771	1.524	2.609	0.009	0.986	6.968
C(SEASON)[T.winter]	-0.7005	1.637	-0.428	0.669	-3.913	2.512
C(WIND_DIR)[T.NW]	5.7812	1.412	4.094	0.000	3.010	8.552
C(WIND_DIR)[T.SE]	1.6281	1.439	1.132	0.258	-1.195	4.451
C(WIND_DIR)[T.SW]	4.6030	1.356	3.394	0.001	1.942	7.264
NO2	-19.3214	84.186	-0.230	0.819	-184.520	145.877
O3	242.0036	49.391	4.900	0.000	145.083	338.924
SO2	3525.4903	620.987	5.677	0.000	2306.920	4744.060
CO	87.1251	5.498	15.846	0.000	76.336	97.914
TEMP	2.2188	0.607	3.653	0.000	1.027	3.411
지면온도	-0.4537	0.230	-1.975	0.049	-0.904	-0.003
평균이슬점온도	-1.8943	0.613	-3.090	0.002	-3.097	-0.691
ATM_PRESS	0.2688	0.102	2.627	0.009	0.068	0.470
WIND	4.0301	0.743	5.425	0.000	2.572	5.488
RAIN	-0.0179	0.044	-0.402	0.688	-0.105	0.069
강수계속시간	-0.6942	0.144	-4.836	0.000	-0.976	-0.413
HUMIDITY	0.6626	0.178	3.732	0.000	0.314	1.011
SNOW	-0.4574	0.940	-0.486	0.627	-2.302	1.388
CLOUD	0.0443	0.184	0.241	0.810	-0.317	0.406
안개계속시간	-2.7682	1.776	-1.559	0.119	-6.253	0.717

모델링 결과

- EDA 5-5절과 다르게, O3가 1증가할 때 PM10이 216증가한다.
- CO가 1증가하면 PM10이 108 증가
- WIND가 증가하면 PM10이 3.8 증가 (가설과 반대다)
- 강수계속시간이 PM10에 미치는 영향은 꽤 미미하다 (-0.49)
- 계절과 풍향에서는
- 계절은 겨울일 때 PM10이 -8로 가장 크게 하락하고, 봄일 때 -1로 가장 높은 값을 가진다.
- 풍향은 동쪽일 때 -9, -7로 하락하고, 서쪽일 때 -1.5, -3으로 조금 더 적게 하락한다.

RMSE의 경우 train/test에 대해 약 11~12 수준이 나왔다.

- PM10의 분포를 다시 살펴보면, 평균은 36.0 / 표준편차는 20.5; Min=3.0 / max=177.0 이다.
- 다중회귀모형의 예측값이 대략 10~13 정도의 오차를 보인다는 말인데, 크다면 크고 작다면 작다 (미세먼지의 좋음,나쁨 기준이 0~30/30~80/80~150 정도니까; 꽤나 자주 미세먼지 좋음/나쁨 등 기준을 벗어난다는 말이다.

Dep. Variable:	PM10	**R-squared:**	0.662
Model:	OLS	**Adj. R-squared:**	0.658
Method:	Least Squares	**F-statistic:**	139.7
Date:	Tue, 31 Aug 2021	**Prob (F-statistic):**	1.85e-160
Time:	22:24:42	**Log-Likelihood:**	-2864.1
No. Observations:	723	**AIC:**	5750.
Df Residuals:	712	**BIC:**	5801.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1729	2.119	-9.992	0.000	-25.333	-17.013
O3	216.6043	50.857	4.259	0.000	116.757	316.451
CO	108.1567	3.637	29.739	0.000	101.017	115.297
WIND	3.8972	0.759	5.136	0.000	2.408	5.387
강수계속시간	-0.4938	0.104	-4.732	0.000	-0.699	-0.289
SEASON_autumn	-7.4018	0.918	-8.060	0.000	-9.205	-5.599
SEASON_spring	-1.0523	1.167	-0.902	0.367	-3.343	1.238
SEASON_summer	-4.6853	1.026	-4.565	0.000	-6.700	-2.670
SEASON_winter	-8.0335	1.257	-6.390	0.000	-10.502	-5.565
WIND_DIR_NE	-9.4838	1.202	-7.890	0.000	-11.844	-7.124
WIND_DIR_NW	-1.5009	0.957	-1.568	0.117	-3.380	0.379
WIND_DIR_SE	-7.0513	1.124	-6.273	0.000	-9.258	-4.844
WIND_DIR_SW	-3.1370	0.975	-3.218	0.001	-5.051	-1.223

5. Decision Tree Regressor

모델링

- 모든 설명변수를 넣고 모델링한 결과, 변수중요도는 오른쪽과 같다.
    - 미세먼지 예측에 일산화탄소의 변수중요도가 약 0.7로 가장 높았고, 그 다음이 평균이슬점온도, SO2, O3 등이었다.

결과

- Train R^2: 0.76
- Test R^2: 0.424
- Test MSE: 169.224
- Test RMSE: 13.009

RMSE만 놓고 봤을 때는 다중회귀모형보다 약간 낮은 성능을 보여줬다.

	feat	imp
3	CO	0.708298
6	평균이슬점온도	0.074551
2	SO2	0.067108
1	O3	0.066642
11	HUMIDITY	0.030064
10	강수계속시간	0.018107
4	TEMP	0.010352
8	WIND	0.007101
7	ATM_PRESS	0.005139
0	NO2	0.005088

Random Forest

- 랜덤포레스트의 변수중요도도 일산화탄소가 가장 높다는 점에서 앞의 두 모델 (다중회귀분석, 의사결정나무)과 비슷했다.

	Feature	Importance
3	CO	0.496876
1	O3	0.068928
2	SO2	0.064089
11	HUMIDITY	0.053806
0	NO2	0.047949
6	평균이슬점온도	0.045983
8	WIND	0.039887
7	ATM_PRESS	0.032724
4	TEMP	0.031448
5	지면온도	0.028569

Gradient Boosting

- 부스팅 기법을 사용한 결과도 변수중요도는 일산화탄소가 가장 높았다.

- 앙상블 기법(Random Forest, Gradient Boosting)의 경우에는 일산화탄소 이후에 오존/습도/아황산가스의 중요도가 높게 나왔다는 점에서 꽤 비슷한 양상을 보여준다.

	Grid Feature	Grid Importance
3	CO	0.568
1	O3	0.079
11	HUMIDITY	0.069
2	SO2	0.062
8	WIND	0.030
5	지면온도	0.026

모델 성능 비교

	Train R^2	Test R^2	Train MSE	Test MSE	Train RMSE	Test RMSE
Multiple Linear	0.662	0.520	161.585	141.126	12.712	11.880
Decision Tree	0.760	0.424	115.002	169.224	10.724	13.009
Random Forest	0.950	0.643	19.874	175.710	4.458	13.256
Gradient Boosting	0.896	0.634	41.476	166.239	6.440	12.893

모델 성능 순위

Test R^2 순위

- Gradient Boosting > Random Forest > Multiple Linear > Decision Tree

Test MSE/RMSE 순위

- Multiple Linear > Gradient Boosting > Decision Tree > Random Forest

앙상블 모델은 상대적으로 train data에 과적합되는 경향을 보였다.

모델 별 유의변수

모델링 결과 가장 유의한 변수는

- 일산화탄소(CO)였다.
- 그 다음으로는 오존(O3), SO2, HUMIDITY, NO, 평균이슬점온도 등이 자주 등장하는 변수다.

기존에 세웠던 가설들과 관련된 변수들(계절, 강수량, 풍속, 풍향 등)이 모델링 과정에서도 유의한 변수들로 도출되고 모델의 설명력에 큰 중요성을 차지했더라면, 기존에 세웠던 가설들을 검정하는 것이 더 쉬웠을 것이다.

- 그러나 EDA/사전학습으로 추측했던 유의변수와 모델링을 통해 도출된 유의변수 간에 차이가 보이면서, 위의 데이터를 직관적으로 이해하기가 더 어려웠다.

7. 피드백 / 개선 방향

데이터 관련

- 확보한 데이터가 '일(day)' 단위인 점이 아쉬웠다.
    - 미세먼지 농도는 시간별로 크게 변동하는 경우가 많기 때문에, 일평균 미세먼지 농도로는 여러 기상/대기오염 데이터와의 상관성을 파악하기 어려울 수 있다.
- 추가적으로 한반도와 가까운 중국 지역의 기상데이터를 확보하면 국내 미세먼지 농도와의 상관관계를 탐색하는 데 좋을 것 같다.
- 수도권 차량 밀집도라던가, 산업시설의 활동량 데이터도 확보한다면 도움이 될 것이다.

분석 관련

- 시계열 분석을 적극적으로 활용하지 못한 점이 아쉽다. 분명 미세먼지 농도가 계절적인 사이클을 띠고 있는데, 그 특징을 분석에 적극 활용하면 더 깊은 인사이트를 얻어낼 수 있을 것 같다.
- 일산화탄소의 설명력이 꽤 높은 것 같은데, 데이터상 일산화탄소 결측치가 많았던 점이 아쉽다.
    - 다른 몇몇 변수들을 이용해 회귀식을 만들어 결측치를 채운 것이기 때문에 일산화탄소와 다른 변수들간 다중공선성이 강화됐을 여지가 있어 보인다.

감사합니다.