머신러닝 기법을 활용한

미세먼지 영향 변수 평가 연구

- PROGRESS REPORT -

김 진 형

2017.08.31

지난 progress 지적 사항

1. 중국 데이터 대표성 문제

2. 대기오염물질 배출량 데이터 검증

3. 결과를 내기

중국 데이터 대표성 문제

베이징(2008~2016 PM2.5)데이터만 쓴 것이 중국을 잘 대표할 수 있는가
- 고비사막과 텐진항 데이터 구하기 어려움
- 상해 데이터 제안

상해(2011~2016 PM2.5)데이터를 베이징 데이터와 같은 방식으로 처리
- 상해와 각 시군구 간의 거리로 min-max 표준화

표준화 방식
- z-score 방식
- 추후 작업하겠음

대기오염물질 배출량 데이터 검증

대분류 변경에 따른 데이터 안정성 문제
- 2007년 배출원 대분류 체계 변경
- 배출량 산정방법의 변화는 꾸준히 있었음 (배출계수, 적용도 등의 변화)
- 데이터 오류

대분류 및 연도별 그래프를 통해 육안으로 판단 (사진 파일 참조)
- 다소 문제가 있는 부분도 있으나 임의로 데이터를 만지기에는 어려움이 있음
- 별다른 수정 없이 그대로 쓰거나 07년 이후 데이터만 쓰는 방법

결과를 내기 – 결과 해석

Decision Tree의 육안 분석

– 결과를 보고 해석, 가장 상위 노드가 제일 중요한 변수이고 분류에 이용되는 노드들이 중요 변수라고 해석

Decision Tree와 Random Forest의 정량 분석

- Classification의 경우 이익지수를 통해서 노드의 impurity가 줄어드는 정도로 변수의 중요성을 비교함
- Regression의 경우 에러의 감소량으로 변수의 중요성을 비교함

결과를 내기 – 실제 적용

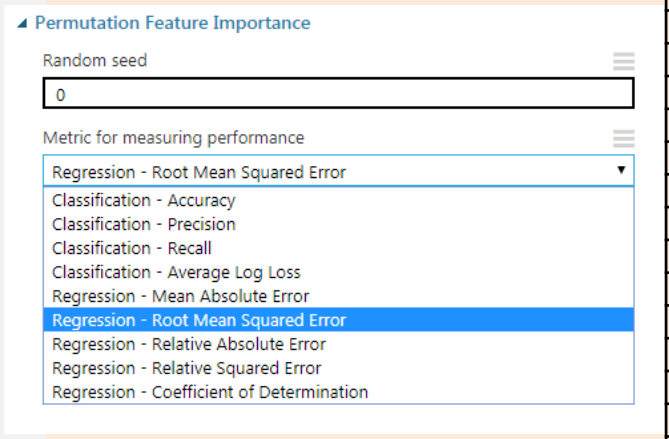
Microsoft Azure (Permutation Feature Importance)

– 한 변수를 제거했을 때, 증가하는 에러를 통해서 변수의 중요도를 평가함

R package (rpart, randomForest)를 이용한 분석

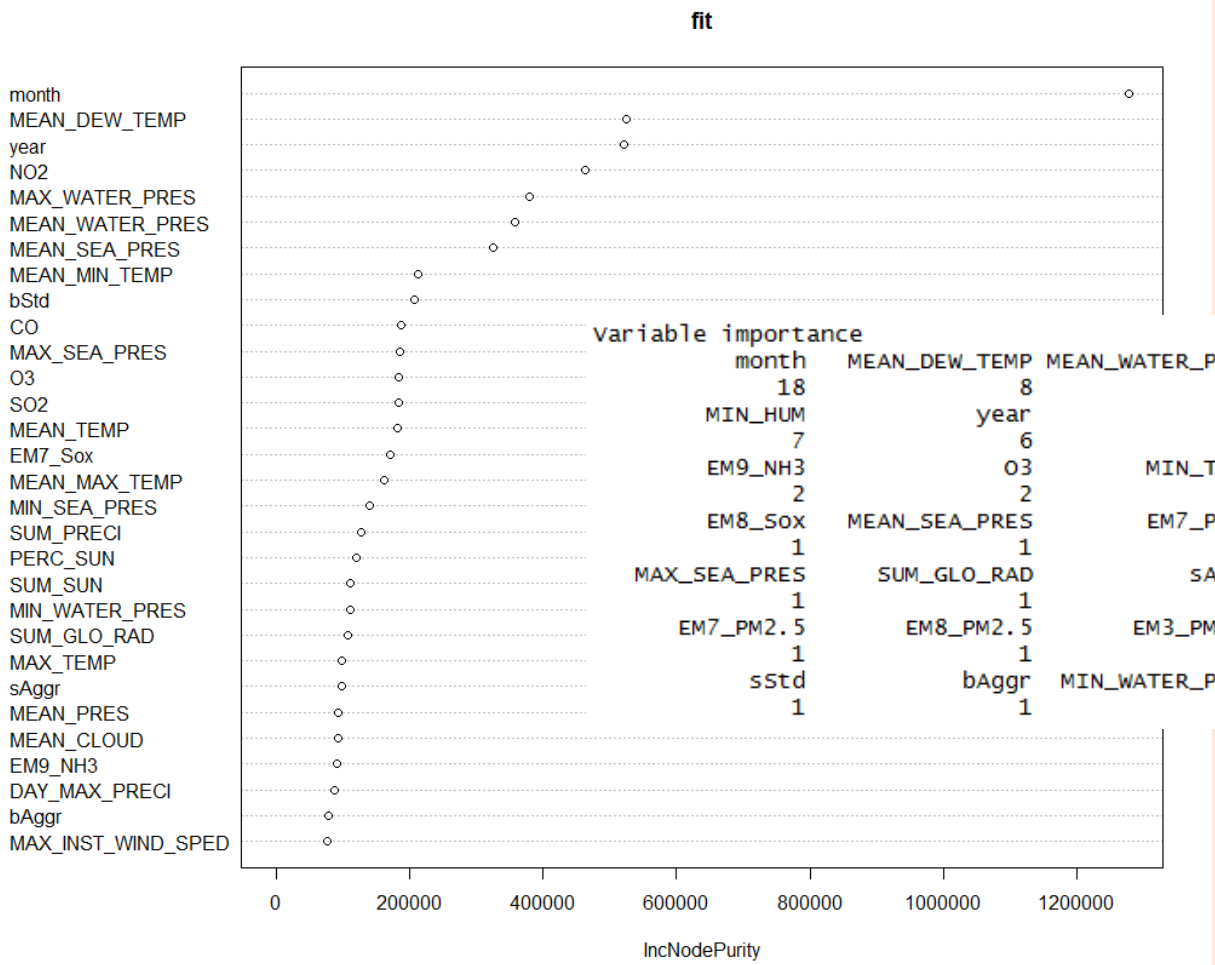
- Classification의 경우 노드의 impurity가 줄어드는 정도로 변수의 중요성을 비교
- Regression의 경우 impurity가 줄어드는 정도를 쓰는 것 같은데 좀 더 봐야 함

결과를 내기 – Azure



1	month	3.22776
2	NO2	1.1682
3	year	0.96217
4	MEAN_SEA_PRES	0.80856
5	MEAN_TEMP	0.71475
6	EM3_PM2.5	0.35397
7	MAX_WATER_PRES	0.3298
8	bAggr	0.32664
9	MIN_SEA_PRES	0.31774
10	bStd	0.31715
11	MEAN_DEW_TEMP	0.27023
12	MEAN_WATER_PRES	0.19476
13	Column 0	0.19072
14	SO2	0.17785
15	EM8_PM2.5	0.17165
16	sAggr	0.16369
17	MIN_TEMP	0.15934
18	MEAN_MIN_TEMP	0.13044
19	EM9_NH3	0.12485
20	MEAN_PRES	0.11079
21	MEAN_CLOUD	0.10815
22	SUM_PRECI	0.10336
23	MEAN_HUM	0.09449
24	MAX_SEA_PRES	0.08891
25	O3	0.08491
26	SUM_GLO_RAD	0.08384
27	MIN_HUM	0.07779
28	PERC_SUN	0.07488
29	SUM_SUN	0.06126
30	MAX_INST_WIND_DIR	0.06028

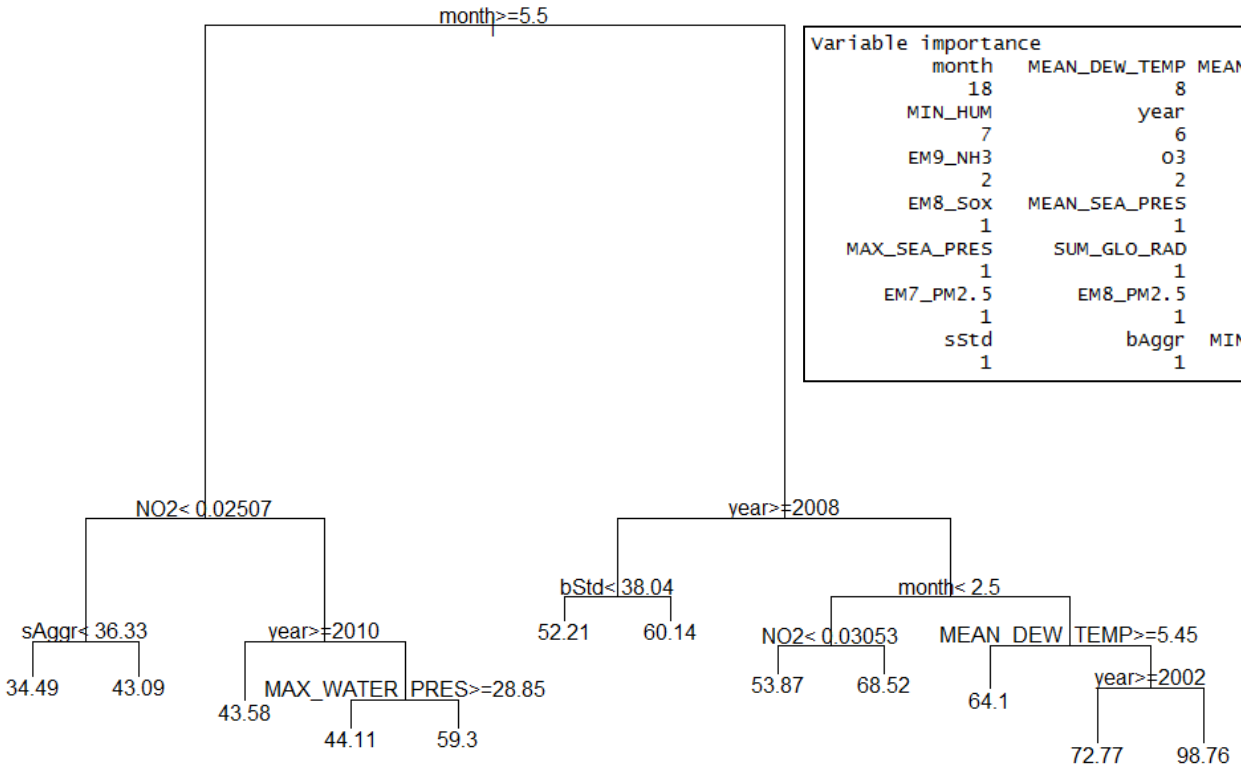
결과를 내기 – randomForest



Variable importance

month	MEAN_DEW_TEMP	MEAN_WATER_PRES	MEAN_MIN_TEMP	MAX_WATER_PRES
18	8	8	8	8
MIN_HUM	year	NO2	EM7_Sox	MEAN_TEMP
7	6	5	2	2
EM9_NH3	O3	MIN_TEMP	MAX_TEMP	MEAN_MAX_TEMP
2	2	2	1	1
EM8_Sox	MEAN_SEA_PRES	EM7_PM10	EM7_TSP	CO
1	1	1	1	1
MAX_SEA_PRES	SUM_GLO_RAD	sAggr	EM2_VOC	EM2_PM2.5
1	1	1	1	1
EM7_PM2.5	EM8_PM2.5	EM3_PM2.5	MEAN_CLOUD	bStd
1	1	1	1	1
sStd	bAggr	MIN_WATER_PRES		
1	1	1		

결과를 내기 – rpart



Variable importance

month	MEAN_DEW_TEMP	MEAN_WATER_PRES	MEAN_MIN_TEMP	MAX_WATER_PRES
18	8	8	8	8
MIN_HUM	year	NO2	EM7_Sox	MEAN_TEMP
7	6	5	2	2
EM9_NH3	O3	MIN_TEMP	MAX_TEMP	MEAN_MAX_TEMP
2	2	2	1	1
EM8_Sox	MEAN_SEA_PRES	EM7_PM10	EM7_TSP	CO
1	1	1	1	1
MAX_SEA_PRES	SUM_GLO_RAD	sAggr	EM2_VOC	EM2_PM2.5
1	1	1	1	1
EM7_PM2.5	EM8_PM2.5	EM3_PM2.5	MEAN_CLOUD	bStd
1	1	1	1	1
sStd	bAggr	MIN_WATER_PRES		
1	1	1		

결과를 내기 - 어려움

해석하는 문제
- 의사결정 나무 단순 해석? -> 어려움
- PFI, VI 등을 활용? -> 정량적 표현이 어려움

Variance가 큼
- Tool마다 결과 다름

향후 계획

Data 쪼개기
- 데이터 구축 연도에 따라 data를 나눔

Data 추가
- 위경도 좌표 추가

문헌분석
- 결과해석 방법 찾기

Thank you!