결과 보고서

소속	부산대학교	
팀 명	2419	
팀 원	권병근, 박민서, 성다솜, 이다인	
제 목	XGBoost + Oversampling + SHAP_Value을 활용한 시정계급 예측 및 유의미한 변수 분석	
과 제	기상특성에 따른 안개 발생 진단	
활용 도구	작업환경	Jupyter Notebook (Anaconda3) Python version: 3.11.5
	라이브러리	Pandas (2.0.3) matplotlib (3.7.2) scikit-learn (1.5.0) imbalanced-learn (0.12.3) xgboost (2.0.3) lightgbm (4.3.0) shap (0.45.1)

Introduction

본 과제의 목표는 주어진 기상관측자료의 특성을 분석하여 시정 계급 구간을 예측하고 안개 발생 진단 모델을 제시하는 것이다. 이 모델은 지표면에서 정상적인 시각의 사람이 목표를 식별할 수 있는 최대 거리인 시정에 따라 총 4단계의 시정 계급으로 분류한다.

본 팀은 시정 계급을 예측하기 위해 다양한 파생 변수를 생성하고, 오버샘플링 기법을 비롯해 여러 머신러닝 모델 앙상블을 통해 일반화 성능을 향상시켰다. 모델 구축 후 각 변수가 결과값에 미치는 영향을 파악하기 위해 SHAP Value 분석 기법을 사용했다. 모델 평가 결과, CSI 점수가 0.105로 검증됐다. 시정계급에 결정적인 영향을 주는 변수는 습도, 강수 유무, 기온 차, 이슬점 등이 있다. 이중에 습도는 낮아질수록 시야가 넓어져 시정계급이 4일 확률이 높아진다. 현 시점에 비가오면 시정계급이 1일 확률이 올라가는데 이는 폭우로 추정할 수 있으며, 24시간 내에 비가 오고 특정 지역에서 시정계급이 2일 확률이 높아 짙은 안개라고 추정할 수 있다. 마지막으로 시정계급 3은 습도와 기온 차 이슬점의 영향을 많이 받으며 21시간내의 강수 유무가 영향을 많이 받아 옅은 안개임을 추정할 수 있다.

Data Preprocessing

- 데이터 정보

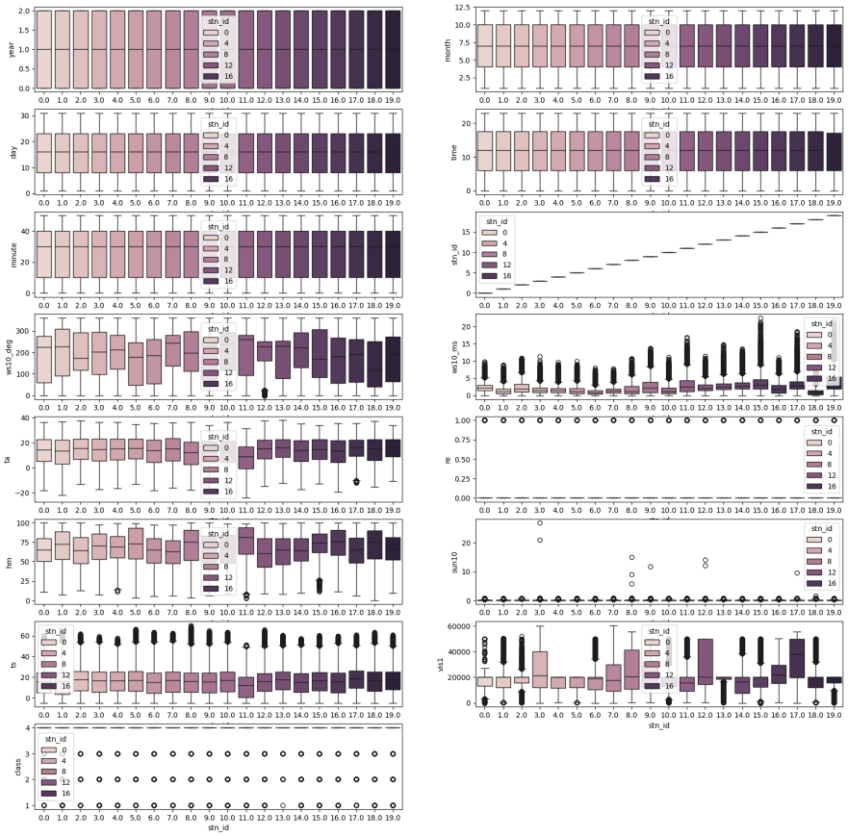
10분 단위의 기상(풍향, 풍속, 기온, 상대습도, 일사량, 지면온도, 시정)의 평균 데이터와 지역특성에 따라 5개의 지역(내륙, 내륙산악, 동해안, 서해안, 남해안)으로 구분한 ASOS 20개 지점 데이터가 있다.

train 데이터는 (3156459, 15), test 데이터는 (262800,14)의 크기를 가지며 컬럼 개수의 차이는 vis1 컬럼의 유무로 발생했다.

컬럼 명	내용	컬럼 명	내용
year	년도	re	강수 유무 (0:무강수, 1:강수)
month	월	hm	1분 평균 상대 습도 10분 주기 (단위: %)
day	일자	sun10	1분 일사량 10분 단위 합계 (단위: MJ)
time	시(0~23)	ts	1분 평균 지면온도 10분 주기 (단위: ℃)
minute	분(10분 단위)	vis1	1분 평균 시정 10분 주기 (단위: m)
stn_id	지점 번호	class	시정 구간
ws10_deg	10분 평균 풍향 (단위:deg)	dew	이슬점 (단위: ℃)
Ws10_ms	10분 평균 풍속 (단위: m/s)	taVSdew	이슬점과 평균 기온 차 (단위: ℃)
ta	1분 평균 기온 10분 주기 (단위: ℃)	tsVSta	평균 기온과 지면 온도 차 (단위: ℃)

- 탐색적 데이터 분석(EDA)

해당 그림은 각 stn_id별 feature의 boxplot을 시각화한 것이다. 이때, stn_id는 0부터 19까지 고유 번호를 부여한 후 시각화를 진행했다. 여기서 특히 주목해 볼 점은 강수 유무를 뜻하는 re가 모든 지역에 대하여 1을 이상치로 가지므로, 비가 온 것이 비가 오지 않은 것에 비하여 매우 적다는 것을 의미한다. 시정 계급을 뜻하는 class에서 4를 제외한 나머지 1,2,3이 모두 이상치로 분류되는 것을 보아 클래스 불균형이 존재한다는 것 또한 알 수 있었고 이후 oversampling을 통해 해결했다. 마지막으로, 지점별 데이터 분포에 차이가 존재한다는 것을 알 수 있다.



- 기본 전처리

1. 컬럼명 단순화

편의를 위해 컬럼 명에 존재하는 'fog_train.'을 모두 생략했다.

2. year 컬럼

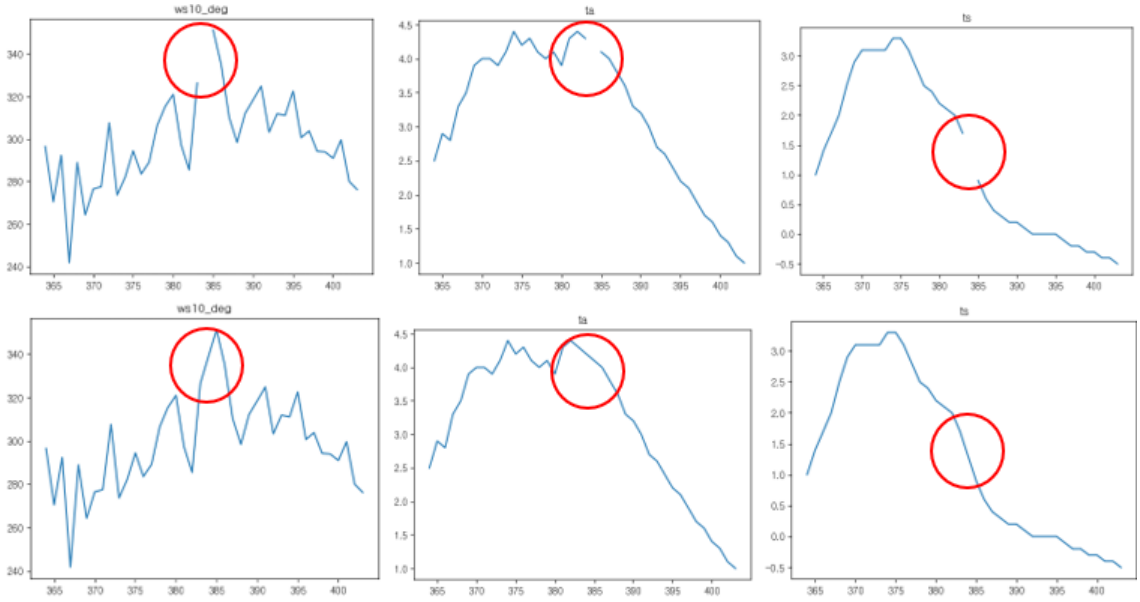
A, B, C 알파벳으로 주어진 year 데이터에 임의의 연도를 부여하여 int형 데이터로 변환했다.

3. stn_id 컬럼 라벨 인코더

알파벳 두 개로 구성된 stn_id는 앞에 위치한 5개의 알파벳에 따라 5개의 지역이 구분되었다고 가정하고 0부터 4까지 5개의 숫자를 부여하여 int형 데이터로 변환해서 학습에 사용했다.

- 결측치 처리

우선 -99.9와 -99 값으로 채워진 결측치를 nan 값으로 변환하여 컬럼별 결측치 개수를 파악했다.



1. float형 결측치

'ws10_deg', 'ws10_ms', 'ta', 'hm', 'sun10', 'ts' 컬럼에 존재하는 float형 타입의 결측치를 linear interpolation을 활용해 처리했다.

2. bool타입 결측치

re 컬럼은 강수 유무를 나타내는 컬럼으로 결측치가 존재할 경우에 비가 오지 않았다고 판단하여 0으로 결측치를 대체했다.

```
기존 학습 데이터 결측치 개수
fog_train.year            0
fog_train.month           0
fog_train.day             0
fog_train.time            0
fog_train.minute          0
fog_train.stn_id          0
fog_train.ws10_deg     5910
fog_train.ws10_ms      5826
fog_train.ta           3867
fog_train.re          15228
fog_train.hm           3616
fog_train.sun10       43910
fog_train.ts          88639
```

```
결측치 처리 후, 학습 데이터 결측치 개수
year        0
month       0
day         0
time        0
minute      0
stn_id      0
ws10_deg    0
ws10_ms     0
ta          0
re          0
hm          0
sun10       0
ts          0
```

3. 타겟 데이터 결측치

vis1과 class 컬럼에 존재하는 결측치는 22516개로 전체 데이터의 0.7%이다. 따라서 제거해도 학습에 별 영향을 주지 않을 것이라 판단해 dropna를 통해 제거 후, 학습에 사용하지 않았다. 결측치 처리 후, train 데이터의 크기는 (3133943,15)로 22516개 감소했다.

- 파생변수 생성

1. 이슬점 변수인 dew 생성

상대 습도와 평균 기온을 활용해 이슬점 파생변수를 생성했다.

$$dew = \left(\frac{hm}{100}\right)^{\frac{1}{8}}(112 + 0.9 + ta) + (0.1 * ta) - 112$$

2. 기온과 이슬점 온도 차 변수인 taVSdew 생성

$$taVSdew = ta - dew$$

3. 평균 기온과 지면온도 차 변수인 tsVSta 생성

$$tsVSta = ts - ta$$

파생변수 생성 이후, train 데이터의 크기는 (3133943,18)로 컬럼이 3개 증가했다.

4. n시간 내 강수 유무(re_3, 18, 21, 24)

시정 계급 1, 2, 3으로 분류할 때, 현재 시점 기준 이전 시간의 강수 유무가 영향을 많이 준다는 점을 고려해 3시간 단위로 24시간 전까지 강수 유무를 나타내는 파생변수를 생성했다. 시간 내에 비가 한 번 이상 왔을 경우 1, 한 번도 오지 않았을 경우를 0으로 나타냈다. 통계테스트를 통해 p-value 0.05 이하여서 95%신뢰구간에 속하는 3, 18, 21, 24시간 이전 강수 유무를 나타내는 파생변수를 활용했다.

- 오버샘플링

Class 값이 4인 데이터에 비해 1,2,3인 데이터가 현저히 적으므로, 데이터 불균형이 발생한다. 4인지 아닌지를 분류하는 모델을 우선적으로 만들기 위해, 1,2,3 클래스는 0으로, 4는 1로 변환을 해주고 1 : 2의 비율로 오버샘플링을 했다. imbalanced-learn 모듈을 이용해 오버샘플링의 여러 방법 중, 가장 좋은 성능을 보여주었던 SMOTE 기법을 선택했다. SMOTE 기법은 최근접 이웃(k-NN) 알고리즘을 활용하여 낮은 비율로 존재하는 클래스의 데이터를 새롭게 생성하는 오버샘플링 기법으로, 이를 활용해 1,2,3을 변환한 값이 0 데이터를 새롭게 생성해 데이터 불균형을 해결했다.

class	오버샘플링 전	오버샘플링 후
1(4)	3,101,809	3,101,809
0(1,2,3)	32,134	1,550,904

모델링

모델링은 총 2단계에 걸쳐 진행되었는데, 4인지 아닌지를 우선적으로 분류하고, 4가 아닌 것들을 다시 1,2,3으로 분류하는 과정을 통해 총 4단계의 시정 계급을 분류하는 모델링을 진행했다. 머신러닝에서 주로 사용되는 XGBoost(eXtreme Gradient Boosting), LightGBM(Light Gradient Boosting Machine), Extra Tree, RandomForest 등의 모델을 활용하여 실험을 진행했다. 그 결과,

성능이 가장 우수했던 XGBoost 모델을 최종적으로 선택했다.

시정 계급이 4인지 아닌지 분류하는 모델은 이진 분류 문제이기 때문에 XGBoost, LightGBM, RandomForest, Extra Tree 모델 모두가 좋은 성능을 보였다. 따라서 4와 4가 아닌 데이터의 예측 비율을 고려해 실제 데이터와 가장 가까웠던 모델인 XGBoost로 선정했다. 앞선 모델을 통해 4가 아니라고 예측된 데이터들은 1, 2, 3으로 재분류하는 모델링을 진행했다. 1, 2, 3을 분류할 때는 데이터 전처리 파트에서 언급했던 3시간, 18시간, 21시간, 24시간 이내 강수 유무를 나타내는 파생변수 4개를 더 추가해 모델 학습을 위한 데이터를 준비했다. XGBoost, LightGBM, RandomForest, Extra Tree를 비롯해 XGBoost와 LightGBM의 Hard Voting을 활용한 앙상블을 모델을 실험해 보았다. 우선 RandomForest와 Extra Tree는 자체적으로 검증한 CSI 점수는 좋았지만, 제출한 결과는 상당히 좋지 않아 과적합된 모델임을 알 수 있었다.1, 2, 3 분류 또한XGBoost가 자체 검증에서 나왔던 CSI 점수와 웹사이트 검증을 통해 나온 CSI 점수 모두 가장 좋게 나왔다. 따라서 두 단계의 XGBoost 모델을 통해 시정 계급을 예측하는 모델을 채택했다.

4 분류 모델	1,2,3 분류 모델	자체 검증 CSI
XGBoost	XGBoost	0.096
XGBoost+LightGBM	XGBoost	0.091
LightGBM	XGBoost	0.055
LightGBM	XGBoost+LightGBM	0.061
LightGBM	LightGBM	0.059
RandomForest	RandomForest	0.382
Extra Tree	Extra Tree	0.360

결론 도출

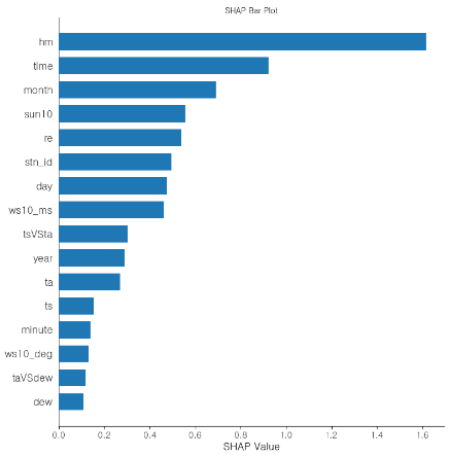
위 모델의 결과를 통해 도출한 결론은 다음과 같다.

1) oversampling과 시정 구간이 시정 계급이 4인 class와 시정 구간이 1, 2, 3인 나머지 class를 먼저 분류하는 것을 통해 클래스 불균형을 문제를 해소할 수 있다. 이는 모델의 결과를 통해 뒷받침할 수 있다.

2) dew, taVSdew, tsVSta, re_3, re_18, re_21, re_24의 파생변수가 모델에 유의미한 영향을 주었다. 위 부분을 확인하기 위해서, 추가적으로 shap value analysis를 진행했다.

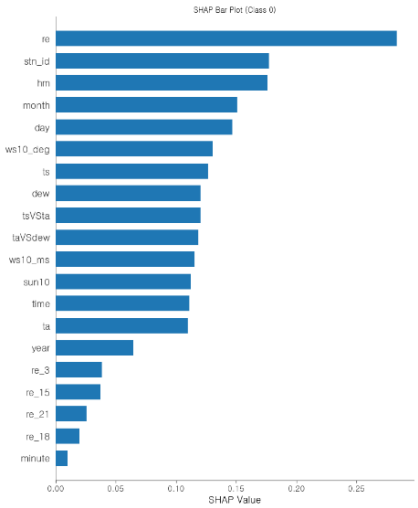
1단계: 시정계급이 4인지 아닌지를 분류하는 모델

위 shap value analysis를 통해서 보았을 때, 시정 계급이 4임을 결정하는데 가장 유의미한 영향을 미치는 것은 hm(습도), time(시)임을 알 수 있다. 시정 계급이 4라는 것은 시정 구간이 매우 커서 시야가 넓다는 것이다. 즉, 습도가 높고, 새벽 시간에 안개를 낄 가능성이 매우 높아지므로 위 두 변수가 클래스를 결정하는데 매우 큰 영향을 미친다는 사실을 쉽게 이해할 수 있다.



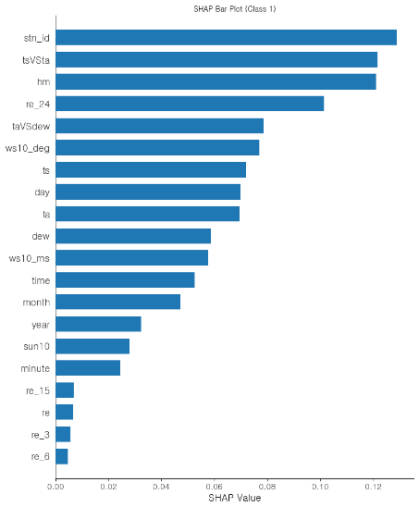
2단계: 1,2,3을 분류하는 모델

해당 모델은 multi-class classification을 수행한다. shap value는 각 클래스를 결정하는 데 특정 feature의 기여도를 수치로 표현한 것이므로, multi-class의 경우 class별로 shap value가 각각 따로 도출된다.



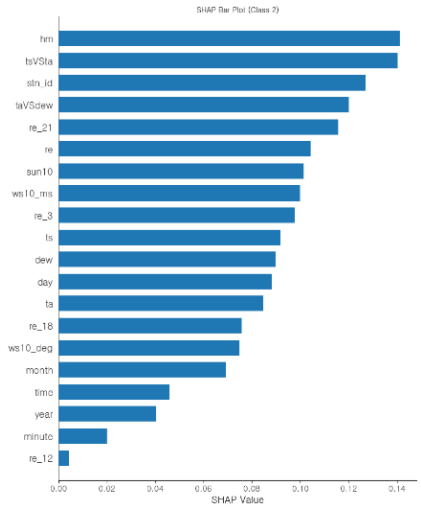
(i) class 0 = 시정계급 1

시정 계급 1을 분류함에 있어서 가장 영향을 많이 미치는 변수는 re, stn_id, hm, month 등이 있다. 이는 안개가 지역의 위치에 따라 영향을 받으며, 강수 유무, 습도, 월 등에 영향을 많이 받는다는 사실로부터 쉽게 유추할 수 있다. 특히 시정 계급이 1인 경우, 시정 구간이 200m 미만으로 매우 짧아 안개의 영향 뿐만 아니라 폭우가 쏟아지는 경우에도 발생할 수 있다. 따라서 강수 유무가 특히 가장 많은 영향을 미친다는 것을 알 수 있다.



(ii) class 1 = 시정계급 2

시정 계급 2인 구간을 분류함에 있어서 가장 영향을 많이 미치는 변수는 stn_id, tsVSta, hm, re_24, taVSdew 이다. 앞 시정 계급 1과 마찬가지로 stn_id, hm 변수가 유의미한 영향을 미친다는 것을 쉽게 알 수 있다. 여기서 주목할 점은 tsVSta, re_24, taVSdew 파생변수로, 지면 온도와 평균 기온의 차, 하루 전의 강수 유무, 평균 기온과 이슬점의 차가 안개 생성에 중요한 변수라는 것을 알 수 있다. 따라서 해당 파생변수들은 모델에 유의미한 영향을 미쳤다



(iii) class 2 = 시정계급 3

시정 계급 3인 구간을 분류함에 있어서 가장 유의미한 영향을 미치는 변수는 hm, tsVSta, stn_id, taVSdew, re_21이다. 이는 시정 계급 2와 마찬가지로 쉽게 알 수 있다.

즉, 위 분석을 통해 생성된 파생 변수가 모델의 결과 및 성능에 긍정적인 영향을 주었다는 사실을 입증할 수 있다.