기상청

기상위성 자료를 활용한 자외선 지수 산출 모델 개발

참 가 번 호	202242	팀 명	기대값

1차 대회- 1과제: 기상위성 자료를 활용한 여름철 자외선 산출기술 개발

1. 분석배경 및 목표

자외선은 태양이 방출하는 광범위한 파장의 빛 에너지의 일부로, 파장이 엑스선보다 길고 가시광선보다 짧은 전자기파이다. 적당한 자외선 노출은 비타민D 합성을 도와주는 등 이로운 점이 있다. 성층권에 존재하는 오존층은 대부분의 해로운 자외선이 지구상의 생명체에 도달하는 것을 막아준다.

그러나 최근에는 자외선을 걸러 주는 오존층의 파괴로 과다한 자외선이 지구로 유입되는 자외선 복사량이 증가하고 있다. 사람들이 오존층을 통과한 자외선을 오랫동안 쬐인다면 피부암, 백내장, 면역결핍증 등 인체에 손상이 온다. 사람들은 자외선을 피하고자 차단제 크림이나 차단 모자, 차단 마스크 등을 사용한다. 따라서 자외선에 대한 정보를 사람들에게 미리 예보하고 자외선으로부터 신체를 보호하는 것은 중요하다. 이에 기상위성 자료를 활용한 여름철 자외선 산출기술 개발하는 것을 목표로 한다.

2. 데이터 정의 및 탐색적 자료분석(EDA)

2020년 1월부터 2021년 12월에 해당하는 자외선 데이터와 국가기상위성센터의 천리안위성 2A호 기본 채널 자료 16종, 태양 천정각 1종, 위성 천정각 1종, 대기외일사량 1종, 지면타입 1종으로 총 20종에 해당하는 데이터를 제공받았다.

2-0) 학습 및 검증데이터의 분리

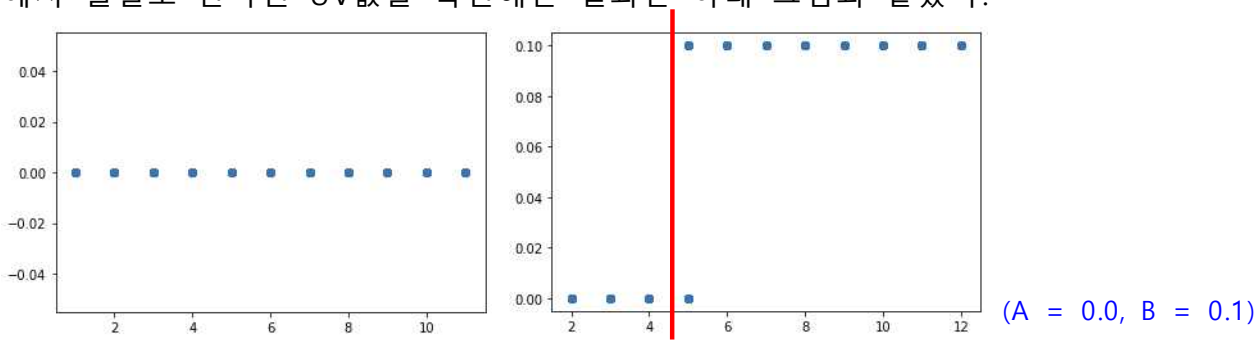
학습모델 설계시 기본적으로 필요한 학습데이터와, 해당 학습모델의 하이퍼파라미터 조정시에 필요한 검증데이터를 나눈 방식은 다음과 같다.

	검증데이터	학습데이터	테스트데이터
날짜	2021년 6월 자외선 관련 데이터	2020년 전체, 2021년 1~5월, 7~12월 UV 데이터	2022년 6월 UV 데이터 (최종 제출용 데이터)
행 개수	781653개	746095개	51855개

[표 1] : 검증데이터와 학습데이터를 나눈 방법

Solarza(태양천정각)에 따른 UV값을 확인해본 결과, UV가 90도 이상일 때, UV값이 [0, 0.1, 0.2] 3가지 값만을 가지는 것을 확인하였다. 따라서 UV가 90도 이상인 상황을 **야간데이터**, 90도 미만인 상황을 **주간데이터**로 나누어 학습모델을 2개 만들었다. 이는

주간의 높은 UV값을 학습한 학습모델의 경향성이 야간데이터에 대한 예측값인 0.0~0.2의 범위를 벗어나는데 기여하는 것을 조금이라도 막기 위함이다. 이후, 2022년 6월 UV 데이터에 대해서도 태양천정각 90도를 기준으로 데이터를 2분할 한 뒤, 2가지의 모델(주간모델, 야간모델)을 각각 적용시킬 것이다. 또한 728710개의 야간데이터의 경우 UV값이 0.1과 0.2가 나오는 상황을 관찰하였는데, 0.1이 15382개, 0.2가 3개인 정황(0.0이 아닌 야간데이터가 약 2%를 차지)을 포착하였고, UV가 0.2인 행은 이상치로 판단하여 삭제하였다. 더하여, 관측지점(STN)별로 UV가 0.0, 0.1, 0.2인 행의 개수를 각각 확인해보았는데, 146 관측소(전라북도 전주시덕진구)에서 UV값이 0.1이 15368개, 0.2가 2개 존재하였다. 즉, 야간데이터에서의 UV값이 0.1, 0.2인 총 데이터 15385개 중, 15370개의 행(야간데이터에서 UV값이 0.1, 0.2인 총 데이터의 약 99.9%)이 146 관측소인 전주에서 발생하였다. 146 관측소에서의 이상치임을 확실히 하기 위해 전주시에서 월별로 관측된 UV값을 확인해본 결과는 아래 그림과 같았다.

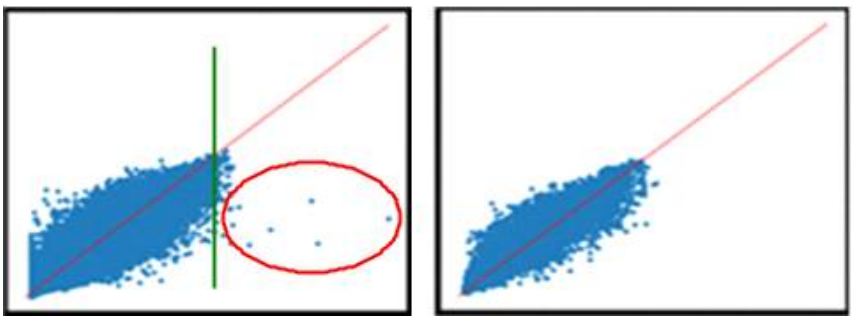


[그림 1] : 2020년 146관측소의 UV값
[그림 2] : 2021년 146관측소의 UV값

그림과 같이 2021년 5월(빨간색 기준선)을 기준으로 월과 상관없이 일정하게 UV값이 0.1로 도출되는 것을 확인하였고, 이러한 상황이 전주시에서만 발생했다는 점에서 이는 관측소 자체의 오류인 특이상황이라 판단하였다. 따라서 야간데이터의 UV값이 0.1, 0.2인 행은 이상치라 결론을 내렸으며 따라서, 해당 행들은 학습데이터인 야간데이터에서 삭제하였다.

2-1) UV 컬럼에 대한 전처리

자외선 값이 –999에 대한 예측은 불가하므로 제공받은 데이터에서 UV가 –999인 행은 모두 학습데이터(야간데이터+주간데이터)에서 제외시켰다. 이후 베이스라인 모델로서 LGBM(Light GBM)을 이용하여 학습데이터와 검증데이터에 대한 예측 UV 값들을 도출하였고, 이에 따른 UV 예측값들과 실제 UV값들 사이의 관계를 아래 2개의 그림으로 표현하였다.



[그림 3] : 학습데이터에 대한 예측 값과 실제 값 사이의 plotting (왼쪽)
[그림 4] : 검증데이터에 대한 예측 값과 실제 값 사이의 plotting (오른쪽)

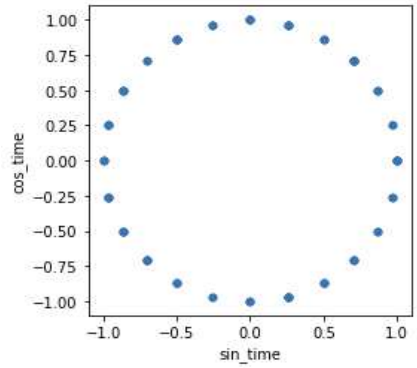
그림 3에서 실제 UV값이 비이상적으로 크게 튀어 오르는 정황을 포착하였고(빨간원), 따라서 UV값이 15이상(그림 3의 초록색 기준선으로부터 오른쪽)인 행을 삭제하였다.

2-2) 16개의 band 컬럼들에 대한 전처리

UV값이 결측치인 행을 모두 삭제한 상태에서 각 컬럼의 결측치(-999)의 개수를 구했더니 band7을 제외한 band1부터 band16 총 15개의 컬럼에서 동일하게 18060의 결측치가 존재하였다(band7에서는 6개 더 많은 18066개의 결측치가 존재했다). 16개의 band 값들을 시계열순으로 관찰한 결과 결측치가 있는 행의 모든 band값들이 결측치인 것을 확인하였다. 전체 데이터 약 150만개에 비해 결측치의 양이 매우 적은 점(약 1.18%)을 근거로 해당 18066개의 행들을 삭제시키고 다음 전처리 식들을 적용시켰다.

2-3) 시간(시계열)관련 컬럼들에 대한 전처리

기존 데이터의 컬럼이었던 'yyyymmdd'와 'hhmm'을 우선 Date 값으로 쓰기 편하도록 'yyyymmddhhmm' 형태의 값을 가진 새로운 파생변수 'YearMonthDayHourMinute'를 만들었다. 이후 이 컬럼을 사용하여 'Year', 'Month', 'Day', 'Hour', 'Min' 5가지의 파생변수를 추가로 만들었으며 기존 'YearMonthDayHourMinute' 컬럼은 삭제하였다. 테스트데이터가 2022년 6월인 점을 감안하여 2020, 2021 값만을 가진 'Year' 컬럼은 학습 사용시에 제외하였다. 월에 따른 UV에 경향성이 존재함을 확인한 바 있으므로, 1~12의 정수형 변수인 'Month' 컬럼은 원핫인코딩을 진행하여 'Month_1'~'Month_12'의 컬럼으로 변환하여 사용하였다. 0~23사이의 정수값을 갖는 'Hour' 컬럼과 같은 경우 단순 수치형 데이터로 학습을 진행 시, 5일 23시와 6일 0시 사이의 간격이 23시간이라고 판단할 수 있다. 이를 방지하기 위해 sin변환과 cos변환을 적용시켜 0~23 총 24가지의 정수형 값을 가지는 1차원 데이터를 sin, cos값을 가지는 2차원 데이터로 변환시켰다. 아래 그림은 'Hour' 변수를 변환한 결과이다. 이 이외의 변수들인 'Day'와 'Min'은 학습에 유의한 영향을 끼치지 않을 것이라 생각하여 학습 시 사용하지 않았다.



[그림 5] : 2차원으로 변환된 'Hour' 변수

2-4) 나머지 컬럼들에 대한 전처리

지면타입에 대한 컬럼인 'landtype'은 0,2,3,4 총 4가지의 값을 가지는 명목형 변수이다. 따라서 'landtype' 컬럼에 대해서 원핫인코딩을 진행하여 'landtype_0'~'landtype_4'의 컬럼으로 변환하였고, 기존 'landtype' 컬럼은 삭제하였다. 관측지점별 수치를 나타내는 컬럼인 'STN'에 대해서는 이미 해당 관측지점의 위도와 경도 값을 저장한 'Lon', 'Lat' 컬럼과 일맥상통하는 면이 있어, 학습시 과적합을 유도할 수 있다고 판단하여 'STN' 컬럼을 삭제하였다.

3. 최종 학습시 사용한 변수 정의

변수명	설명
Lon	관측 지점의 경도
Lat	관측 지점의 위도
band1	파랑 가시밴드
band2	초록 가시밴드
band3	빨강 가시밴드
band4	식생 가시밴드
band5	눈/얼음 채널
band6	권운 밴드
band7	야간안개/하층운 밴드
band8	상층 수증기 밴드
band9	중층 수증기 밴드
band10	하층 수증기 밴드
band11	구름상 밴드
band12	오론 밴드
band13	대기창 밴드
band14	깨끗한 대기창 밴드
band15	오염된 대기창 밴드

변수명	설명
band16	이산화탄소 밴드
solarza	태양천정각
sateza	위성천정각
esr	대기외 일사량
height	관측 지점의 고도
sin_time	시간에 대한 sin 변환
cos_time	시간에 따른 cos 변환
landtype_0~4	지면타입에 대한 원핫인코딩
Month_1~12	월에 따른 원핫인코딩

3. 모델링

3-1) 야간 모델

2-0)에서 설명한 바와 같이 야간 모델의 UV값의 분포는 다음과 같다.

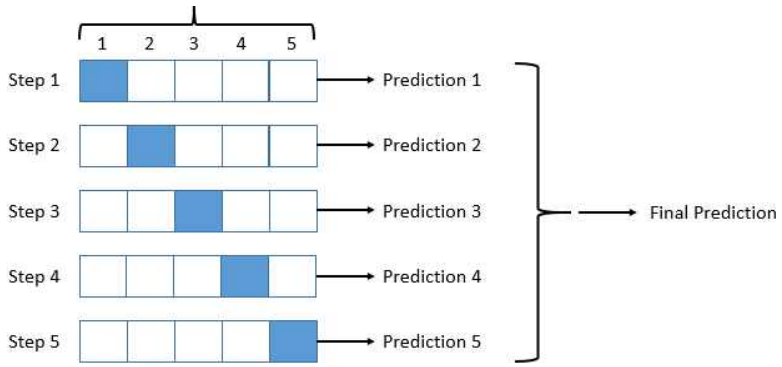
UV값	개수	비율(%)
0.0	728710	97.9
0.1	15382	2.1
0.2	3	0

[표 2] : 야간데이터에서 각 value들의 비율

해당 UV값 중 0.1과 0.2는 STN이 146인 전주 관측소에서 발생한 특이값이라 판단하였다. 따라서 이들을 학습시 제외하였고 그 결과, 야간데이터의 UV가 0으로 모두 동일하였기에 어떠한 모델링 기법도 사용하지 않고 UV값을 0으로 예측하였다.

3-2) 주간 모델

주간학습데이터에 대해서는 단일모델인 LGBM을 사용하였고 Out-of-Fold방식으로 최종 예측결과를 도출하였다. 총 5개의 Fold를 사용하였으며 각 Fold에서의 예측결과인 Prediction 1~5의 평균값을 최종 예측 UV값으로 하였다.



[그림 6] : 5 fold OOF방식에 대한 설명

4. 활용 방안 및 기대효과

현재 기상청은 서울, 포항, 목포, 강릉, 고산, 안면도 등 전국 7곳에 자외선 측정기를 배치해 관측된 자외선 복사량에 특정 파장에 대한 가중치를 곱하고 구름, 대기상태, 고도 자료와 결합해 자외선 지수를 예보한다. 이번 프로젝트를 위해 받은 훈련 데이터에서는 53207개의 UV값이 결측값이었다. 이는 자외선 측정기의 오류이거나 어떤 다른 이유에서 자외선을 측정하지 못한 것으로 파악된다. 이번 공모안에서 제안된 모델은 자외선 복사량의 유무와 관계없이 다른 독립변수를 통해 자외선 지수를 예측하므로 위와 같이 자외선 복사량을 관측할 수 없을 때 사용될 수 있을 것이다. 또한 위도 경도 컬럼과 매우 종속적인 성질을 가졌던 컬럼인 STN 자체를 학습과정에서 삭제하였기 때문에 새로운 위치에 관측소가 들어선다고 하더라도 이전까지 학습한 관측소들의 위도, 경도 값으로 비교적 좋은 성능의 예측을 제공할 수 있을 것이라 기대한다.