기상청

[날씨 빅데이터 콘테스트] 자외선 예측 (과제 1-1)

참 가 번 호	220145	팀 명	날씨맑음

1. 서론

햇빛은 파장에 따라 자외선, 적외선, 가시광선으로 이루어져있다. 이 중 자외선은 파장의 길이의 따라 자외선A(320~400nm), 자외선B(280~320nm), 자외선C(100~280nm)로 구분된다. 자외선A, B, C 중 파장이 가장 짧은 자외선C는 오존층에서 흡수되고, 파장이 가장 긴 자외선A와 자외선B만 지표면에 도달한다. 자외선은 인체 건강에 필요한 비타민D를 합성하는데 중요한 역할을 하며, 비타민D는 근골격계와 뼈 성장에 도움을 주는 인과 칼슘의 흡수를 돕는다. 따라서 자외선에 대한 적절한 노출은 건강에 도움이 된다. 그러나 최근 환경오염으로 인한 오존층의 파괴로 인하여 지표면에 도달하는 자외선이 많아지면서 홍반(화상), 피부암, 백내장 등을 유발하며 인체에 유해한 영향을 준다.

자외선은 기후 변화의 영향을 받는다. 태양고도가 높은 지역이 낮은 지역보다 자외선이 강하다. 오존층은 자외선 복사를 흡수하여 지구의 기후 조절에 중요한 역할을 한다. 그러나 오존층이 파괴되면 오존이 줄어들어 자외선이 강해진다. 구름이 없이 맑은 날도 자외선이 강하게 나타난다. 에어로졸은 공기 중에 떠 있는 작은 입자로 빛을 반사하고 흡수한다. 에어로졸이 늘어나면 흡수되는 빛이 많아지며 자외선이 약해진다. 지표면의 반사율도 자외선에 영향을 준다. 지표면의 반사율은 지면의 상태에 따라 다르지만 반사율이 높아지면 자외선도 높게 나타난다. 해발고도가 높은 지역도 낮은 지역에 비하여 자외선이 강하다.

기상청에서 WMO(세계 기상 기구)에서 제안하는 가이드라인을 활용하여 지표면에 도달하는 자외선A와 자외선B를 바탕으로 자외선 지수를 산출하여 제공한다. 또한 자외선 지수에 따른 단계별 행동 요령을 통하여 일상생활에 도움을 얻을 수 있다. 본 프로젝트는 트리계열의 머신러닝 모델들을 통해 자외선 지수를 예측해보고, 성능 비교를 위해 RMSE(Root-mean-square error)값을 산출해 보고자 한다.

2. 본론

2-1. 데이터 소개

자외선 산출 기술 개발에 사용할 데이터는 기상청의 2번째 위성인 천리안위성 2A호에서 제공하는 기상위성데이터와 기상자료개방포털에서 얻은 기상데이터를 이용했다.

천리안위성 2A호는 가시채널(VIS), 근적외채널(NIR), 단파적외채널(SWIR), 수증기채널(WV), 적외채널(IR) 각각에서 제공하는 총 16개의 채널과 태양 천정각, 위성 천정각, 대기외 일사량, 지면타입 등의 정보를 제공한다.

기상위성데이터에서 제공하는 데이터의 시점과 일치하는 지역의 기온, 풍속, 풍향(16방위), 습도, 증기압, 이슬점온도, 해면기압, 지면온도를 포함한 기상데이터를 이용하여 자외선 산출 기술 개발에 이용했다. 기상위성데이터와 천리안위성 2A호의 데이터의 지점 번호에 차이가 있기에 가장 가까운 지역의 기상데이터로 대체하였다.

	기존	대체
1	제주 황해(13)	고산(185)
2	태안(132)	보령(235)

표 1. 기상위성데이터 대체 지역

2-2. 전처리

이상치나 결측치를 정제함으로써 모델링할 수 있는 변수로 만든다. 제공된 데이터에 결측치는 존재

하지 않았으나 –999로 코딩된 이상치가 uv, band1~band16 총 17개 변수에 각각 다수 존재했다. 이상치 처리의 방법으로 이상치를 모두 제거하거나, 이상치를 적절한 값으로 대체하는 방법이 있다.

그러나 히스토그램을 그려 uv, band1~band19의 –999 존재 패턴을 알아본 결과, 이상치 –999가 전범위에 골고루 퍼져있지 않았다. 따라서 전자의 방법으로 이상치를 처리하면 정보의 손실이 생길 것이라 판단하여, 후자의 방법을 택하였다. -999를 모두 결측치로 바꾸고, 결측치 대체방법인 KNN과 MICE를 사용하여 적절한 값으로 대체하였다.

1) KNN(K-Nearest Neighbor) imputation

근처 k개의 값을 참고하여 결측치를 대체하는 방법이다. k는 5, 7, 10을 사용하였다.

2) MICE(Multiple Imputation by Chained Equation) imputation

결측값을 제외한 다른 데이터로 예측한 값을 채워 넣는 방법이다. mice()함수의 method 옵션에는 random forest를 사용하였다.

2-3. 탐색적 데이터 분석(EDA)

모델을 적합하기 전에, 탐색적 데이터 분석(Exploratory Data Analysis,) 단계를 통하여 분석하게 될 변수를 다각도로 살펴보는 과정이 필요하다.

1) 기초통계량 확인

summary() 함수를 사용하여 각 변수의 최대, 최소 및 사분위수 값을 알 수 있다. 결측치가 없다는 것을 확인하였고, band끼리 단위가 다르지 않으므로 표준화 작업은 진행하지 않았다.

2) 변수 분포 확인

변수의 분포를 확인하고자 hist()함수를 사용하여 히스토그램을 그렸다. 그 결과, uv, band1~band6 변수가 right-skewed 경향을 보이는 것을 확인하였다.

3) 반응변수와 설명변수 관계 탐색

- date에 따른 uv 변화

as.Date()함수를 사용하여 date를 날짜형태로 변환하고 그래프를 그렸다. 날짜(date)에 따른 자외선 지수(uv)의 변화가 뚜렷하다. 회색 세로선은 각

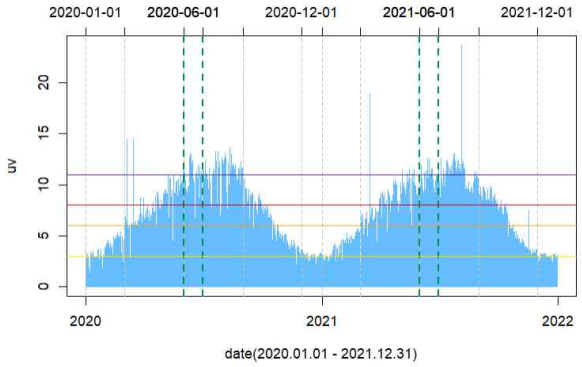


그림 1. 시간에 따른 자외선지수 변화

연도의 1월1일, 3월1일, 6월1일, 9월1일, 12월1일을 나타낸다. 녹색 세로선은 각 연도의 6월 구간을 나타낸다. 계절과 월별로 자외선지수가 달라짐을 알 수 있다. 6월~9월은 매우 높음(빨간색 선)~위험(보라색 선) 단계를 보이며, 9월 이후로는 점차 자외선 지수가 내려간다. 또한 각 연도의 날짜에 따른 자외선지수 추세는 유사하다. 따라서 2020년 3월, 2021년 3,7월의 눈에 띄는 이상치를 제거해준다.

- hhnn, stn에 따른 uv 변화

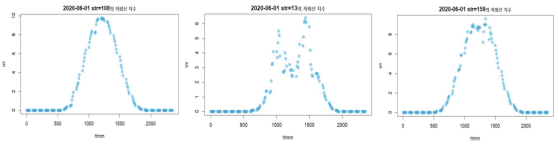
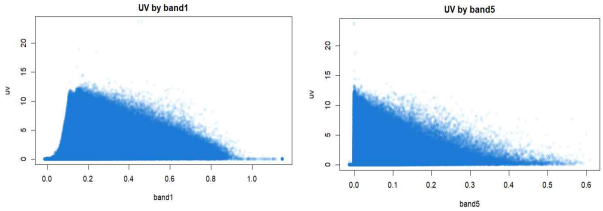


그림 2. 지점 별 자외선지수 변화

왼쪽부터 차례대로 날짜를 하루로 고정(2020년 6월 1일로 고정)했을 때 stn=108,13,159의 시간대별 자외선지수 변화이다. 대체로 정오시간에 자외선 지수가 가장 높다. 그러나 지점번호(stn)과 hhnn(시간/분)에 따라 자외선 지수가 달라지는 것을 보아 지역과 시간대를 예측모델 적합 시 사용할 필요성이 있다.

- 항목별 band에 따른 uv 변화



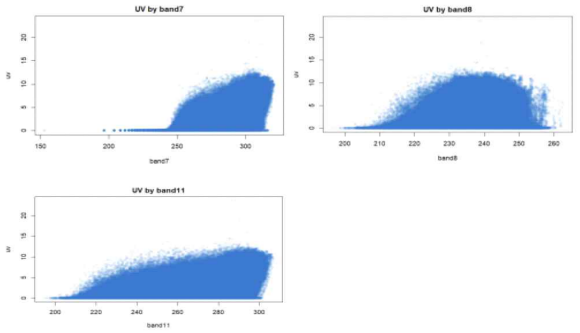


그림 3. 각 밴드 별 자외선지수

각 항목(가시채널(band1~4), 근적외채널(band5~6), 단파적외채널(band7), 수증기채널(band8~10), 적외채널(band11~16))에서 band를 하나씩 뽑아 uv와의 산점도를 그려보았다. 각 채널 별 band와 uv의 관계는 서로 다른 양상을 보인다. 또한 band와 uv와의 관계식이 선형모형만으로 표현되기는 어려워 보인다.

- band, landtype에 따른 uv 변화

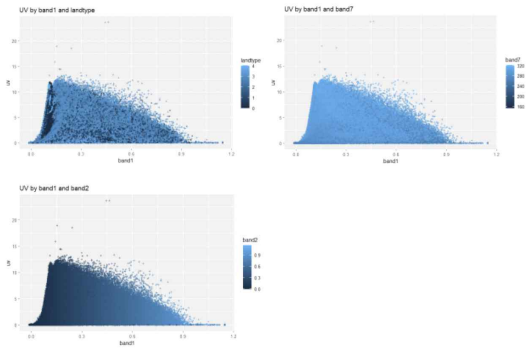


그림 4. 지면 타입 별 자외선지수

왼쪽부터 band1과 landtype에 따른 uv 변화, band1과 band7에 따른 uv 변화, band1과 band2에 따른 uv 변화이다. band1와 landtype, band1과 band7의 상관관계는 낮아 보이지만, 같은 채널끼리인 band1과 band2의 상관관계는 높아 보인다.

2-4. 피처 엔지니어링

본 장에서는 자외선 지수에 영향을 미치는 여러 요인들을 파악함으로써 변수를 추가하거나 적절하게 처리하는 과정을 거쳤다.

1) 태양고도(solaraa)

가덕현 외 2인(2022)에 따르면, 자외선의 세기는 태양 위치의 영향을 받는다고 보고하고 있다. 태양의 위치는 태양고도, 태양 천정각으로 표현할 수 있다. 본 프로젝트의 데이터에서 'solarza' 변수는 태양천정각을 의미하며, 해당 변수를 이용하여 태양고도를 구할 수 있다. 본 프로젝트는 90°에서 태양천정각(solarza)를 차감한 값을 태양고도(solaraa)로 하여 변수를 추가하였다.

$$\text{태양고도}(solaraa) = 90^\circ - \text{태양천정각}(solarza)$$

2) band 선별

두 번째로 2020년 1월부터 2021년 12월 기간 동안 기상위성 관측 데이터의 여러 채널들을 시계열 그래프를 통해 확인하였다. 각 채널은 계절적 추세를 보였으며, 비슷한 패턴을 지닌 채널들을 확인할 수 있었다. 특히 가시채널로 구분되는 band1~band4(그룹 1)는 상당히 유사한 패턴을 보였다. 따라서 해당 채널들은 자외선 지수에 비슷한 방식으로 영향을 줄 것으로 예상하였다.

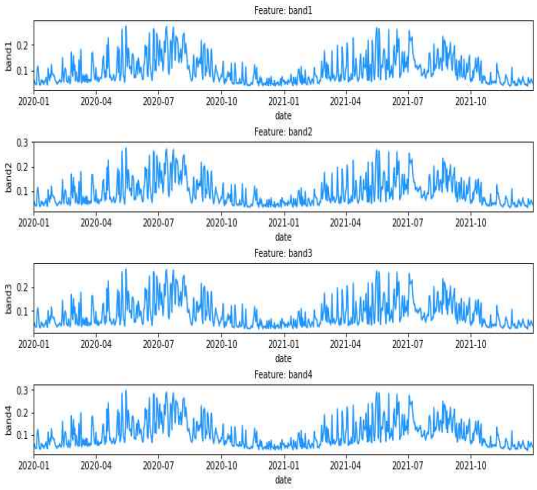


그림 5. band1~band4(그룹 1)의 자외선지수

근적외채널인 band5와 band6, 그리고 단파적외채널인 band7은 시간의 흐름에 따라 각각 고유한 패턴을 지닌 채널로 확인된다. (그룹 2)

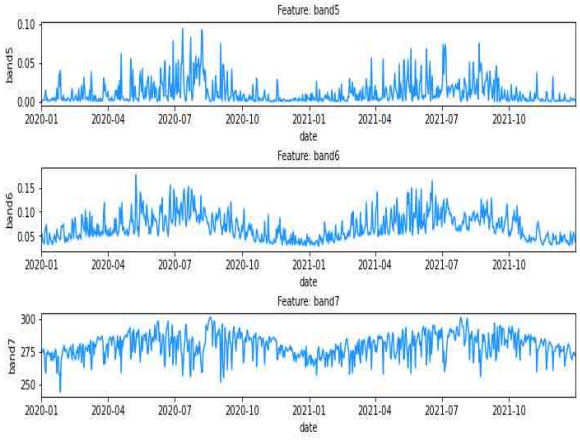


그림 6. band5~band7(그룹 2)의 자외선지수

수증기 채널로 분류되는 band8~band10(그룹 3), 그리고 적외채널로 구분되는 band11~band16(그룹 4)은 모두 유사한 패턴을 보인다.

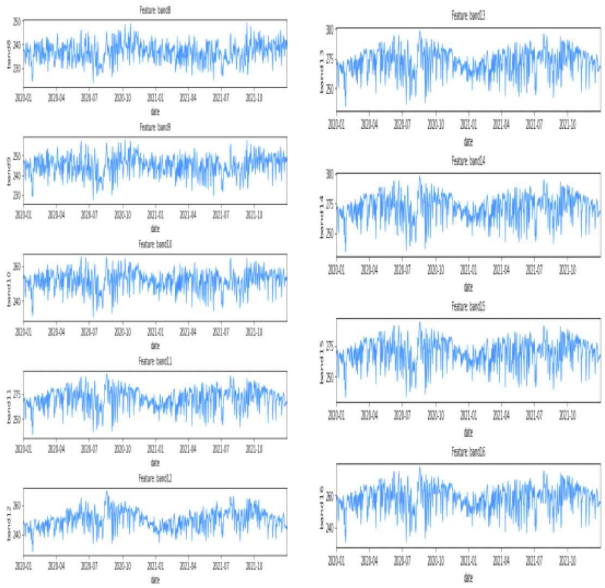


그림 7. band8~band16(그룹 3&4)의 자외선지수

이와 같은 과정을 통해 고유한 특성을 지닌 band들을 선별하여 자외선 지수 예측에 이용하고자 하였다. 예상한 것과 같이 그룹 1(band1~4)과 그룹 2(band5~7)은 모두 사용하는 것이, 그룹 3(band8~10)은 모두 사용하지 않는 것이 정확한 예측에 도움이 되었다. 반면, 예상과 다르게 그룹 4(band11~16)에서 구름상밴드(band 11)의 효과는 눈에 띄지 않았다.

3) 코사인(cos) 변환

마지막으로 월(month)과 일(day) 변수를 코사인(cosine) 변환하였다. 이러한 변수들은 모두 연속적으로 증가하는 숫자로 인식된다. 월(month)의 경우에는 1월부터 12월이 숫자로 표현되기 때문에 머신러닝 모델은 이를 연속적으로 증가하는 숫자로 인식한다. 즉 2020년 12월과 2021년 1월을 11의 차이로 인식하지만, 실제 차이는 1을 의미한다. 이와 같은 데이터를 그대로 머신러닝 모델에 학습하게 되면 연속적인 주기성을 제대로 학습하지 못하기 때문에 코사인(cosine) 변환을 통해 데이터가 연속성을 가질 수 있도록 표현하였다.

(검은색 점: 변환 전 / 파란색 선: 변환 후)

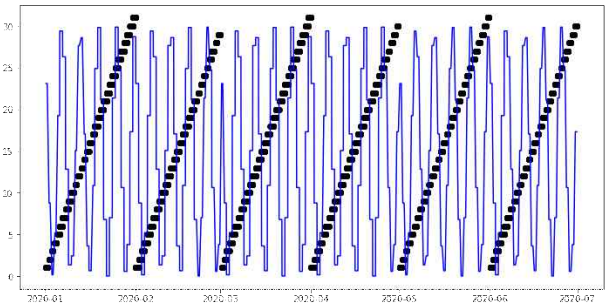


그림 8. 자외선지수의 코사인(cosine) 변환 전/후

2-5. 모델링

1) 모델 선정 기준

모델의 설명 가능성을 중요시하여 블랙박스(Black-Box) 구조를 가지는 딥러닝 알고리즘 대신 머신러닝 알고리즘을 우선적으로 고려한다. 블랙박스란, 모델이 어떠한 근거로 예측 값을 출력했는지 알 수 없는 구조를 의미한다. 반면, 모델의 예측 값을 사람이 이해할 수 있도록 설명하는 방법을 XAI(eXplainable Artificial Intelligence)라고 한다. 설명 가능한 모델의 가장 큰 강점은 신뢰성이다. 인공지능의 발전과 함께 모델의 결과에 대한 사용자 및 사회의 우려도 함께 증가하고 있다. 이러한 상황에서 XAI가 비전문가와 인공지능 모델 사이의 징검다리 역할을 해줄 수 있다. 따라서 변수 중요도(Feature Importance), SHAP 등의 XAI기법을 적용할 수 있는 머신러닝 모델을 선정하였다. 그중에서도 보다 직관적인 형태를 가지는 Random Forest, LightGBM, XGBoost 등의 트리 계열의 모델을 선택했다.

2) 모델 구축 과정

미래 정보를 반영하는 데이터 리키지(Data Leakage)를 하지 않도록 Nested Cross-Validation 방식을 적용했다. 또한, 테스트(Test) 데이터의 기간이 2020.06인 점을 고려하여, 모델 훈련 시 훈련(Train) 데이터와 검증(Valid)데이터의 기간을 다음과 같이 설정하였다.

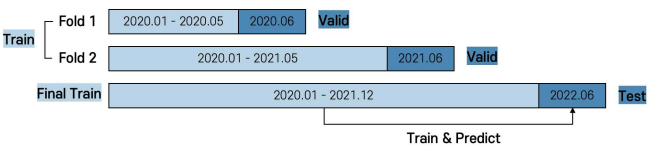


그림 9. Nested-Cross Validation

최종 변수 선택은 두 차례에 거쳐 진행했다. 먼저 탐색적 데이터 분석(EDA)과 피처 엔지니어링

단계를 거쳐 자외선 지수에 영향을 미치는 변수를 선별했다. 선별 후에는 다양한 트리 계열 모델을 학습하였고, 각 모델의 피처 중요도를 확인하여 모델들이 공통적으로 '중요하다/중요하지 않다'고 판단하는 변수를 우선적으로 '고려/제외' 하였다.

월	일	시간	지점번호
파랑 가시밴드	초록 가시밴드	빨강 가시밴드	눈얼음 채널
권운밴드	야간안개/ 하층운밴드	오존밴드	대기창 밴드
깨끗한 대기창 밴드	오염된 대기창 밴드	이산화탄소 밴드	태양 천정각
위성 천정각	대기외 일사량	관측고도	지면타입
풍향	습도	지면온도	태양고도

표 2. 최종 사용 변수

위와 같은 환경으로 모델을 학습한 후, 최종 모델의 예측 값이 음수 형태인 경우 후처리를 진행했다. 자외선 지수는 항상 0 이상의 값을 가지기 때문에 0으로 값을 대체하는 것이 실제 값에 가까울 것이기 때문이다.

2-6. 성능 비교

표 1은 트리계열의 머신러닝 모델인 Random Forest, LightGBM, XGBoost에 대해 학습을 진행하고 성능을 나타낸 것이다. 우선 세 가지의 모델 모두 학습 데이터에서의 성능에 비해 검증 데이터의 성능이 조금씩 감소하는 것을 확인할 수 있다. 이때, LightGBM의 두 번째 폴드에 주목하면 다른 모델보다 학습과 검증의 성능 차이가 작은 것을 확인할 수 있다. 이를 통해 LightGBM이 다른 모델에 비해 일반화된 모델이라 판단한다.

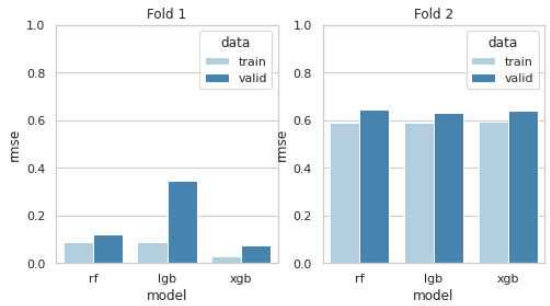


그림 10. 모델 별 성능 비교

Fold	Fold 1		Fold 2	
Model	Train	Valid	Train	Valid
R.F.	0.088	0.591	0.120	0.644
LightGBM	0.164	0.574	0.346	0.631
XGBoost	0.027	0.595	0.077	0.640

표 3. 모델 별 성능 비교

따라서, 최종 모델은 2020.01~2021.12를 학습 데이터로 사용하였고 검증 데이터(2022.06)에 대해 예측한 RMSE 값은 0.604573이다.

2-7. 모델 결과 해석

변수 중요도(Feature Importance)와 SHAP로 모델의 설명력을 해석한다.

1) 변수 중요도(Feature Importance)

변수 중요도는 데이터가 변수 노드에 의해 분할될 때 불순도(Impurity)를 낮추는 변수가 중요하다고 판단한다. 아래 표 4와 표 5는 각 모델 별로 변수 중요도가 높은 상위 5개의 변수와 낮은 하위 5개의 변수를 표로 나타낸 것이다. 모델들이 공통적으로 예측에 중요하다고 판단하는 변수는 지면온도, 태양천정각(solarza), 가시밴드(band)임을 확인할 수 있다. 또한, 예측 시 비교적 중요하지 않다고 판단하는 변수는 지면타입(landtype)과 관련된 변수이다. 이를 통해 향후 자외선 예측 모델을 구축할 때 도움이 될 것이라 기대된다. (단, 각 모델 별 변수 중요도가 가장 큰 값을 1로 했을 때의 수치임.)

R.F.	solarza		solaraa
	1		0.182
	band2	band3	지면온도
	0.102	0.0416	0.0254
LightGBM	band12		hhnn
	1		0.878
	지면온도	band1	습도
	0.634	0.618	0.57
XGBoost	solarza		band2
	1		0.134
	stn_13	band3	지면온도
	0.067	0.049	0.028

표 4. 모델별 중요도가 높은 상위 5개의 변수

R.F.	landtype_3		stn_4
	0		0.0000056
	stn_11	landtype_1	stn_8
	0.000021	0.000022	0.000029
LightGBM	landtype_3		landtype_1
	0		0.00104
	stn_5	stn_7	stn_11
	0.00343	0.0079	0.00819
XGBoost	stn_6		landtype_3
	0		0.00168
	landtype_2	stn_8	stn_11
	0.00188	0.00189	0.00222

표 5. 모델별 중요도가 낮은 상위 5개의 변수

2) SHAP

변수 중요도를 통해 자외선 예측 시 중요한 변수와 비교적 덜 중요한 변수를 알아봤다면, SHAP를 통해서는 해당 변수의 값과 자외선 지수의 관계를 확인해보고자 한다. 아래의 SHAP 그래프에서 SHAP value의 가로 분포 길이가 클수록 영향력 있는 변수라고 판단한다. 예를 들어 태양천정각(solarza), 태양고도(solaraa)의 경우 다른 변수에 비해 눈에 띄게 큰 영향력을 가지는 것을 확인할 수 있다. 태양천정각(solarza)의 값이 작을수록, 그리고 태양고도(solaraa)의 값이 클수록 자외선 지수가 높은 경향을 보인다고 해석할 수 있다.

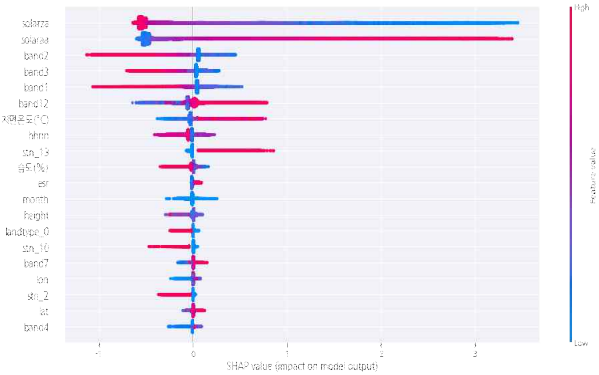


그림 11. SHAP 그래프

3. 결론

본 프로젝트는 여름철 자외선 지수 예측을 위해 자외선 지수에 영향을 미친다고 판단되는 변수들을 통계적 관점에서 분석하였으며, 성능 비교를 통해 LightGBM을 자외선 지수 예측의 최종 모델로 선택하였다. 특히, 자외선 지수에 영향을 미치는 요인들을 통계량에 기반하여 파악하였고, 각 요인들의 히스토그램을 통해 요인별 특성들을 확인할 수 있었다. 또한, 자외선 지수를 예측하기에 앞서 산점도를 통해 자외선 지수와 시간, 지점, 지면타입, 채널, 등 여러 예측 요인들 간의 관계도 파악하였다. 이외에도 채널(band)별 자외선 지수 추이를 시계열 그래프를 통해 확인함으로써 비슷한 패턴을 보이는 것으로 판단되는 채널들을 파악할 수 있었다.

본 프로젝트의 향후 활용 방안은 크게 2가지이다. 첫째, 세분화된 단계별 자외선 지수 서비스를 제공할 수 있다. 기상청은 웹 사이트 '날씨누리'와 모바일 앱 '날씨 알리미'를 통해 현재 지역의 자외선 지수를 단계(낮음, 보통, 높음, 매우 높음, 위험)별로 구분하여 제공한다. 그러나 여름철에는 세분화된 자외선 단계의 필요성이 요구된다. 여름철의 자외선 지수는 다른 계절에 비해 상대적으로 높으며, 자외선 지수의 범위도 넓기 때문이다. 이를 바탕으로 세분화된 단계에 기반하여 인체에 유해한 여름철 자외선에 대비할 수 있을 것으로 기대된다.

둘째, 기상청의 자외선 지수 예측에 관한 의사결정에 도움이 될 것으로 기대된다. 기존 방식과 더불어 머신러닝 모델에 기반한 데이터 분석 결과를 이용한다면, 여름철 자외선 지수의 예측력이 상승할 것으로 보인다.

본 프로젝트는 자외선 지수 예측을 위해 트리 계열의 Random Forest, LightGBM, XGBoost 등의 모델들을 고려하였다. 이외에도 최신의 딥러닝 기술들을 이용하여 분석한다면 본 프로젝트의 한계를 뛰어넘고 더 의미 있는 다양한 정보들을 얻을 수 있을 것이다.

참고문헌

김예진, 남개원. 2021. 각 국가별 자외선차단지수 측정법의 비교 분석 및 자외선차단지수에 영향을 주는 요인들. J. Soc. Cosmet. Sci. Korea Vol. 47, 3, 193-203.

박대환. 2020. 심층신경망을 이용하여 자연광에서 자외선지수를 추정하는 모델 개발. 석사학위논문, 공주대학교 대학원.

이윤곤, 김준, 조희구, 최병철, 김지영, 박일수. 2006. 오존전량, 자외선지수 예측모델의 평가 및 개선. 한국기상학회지, 108-109.

이하나, 김준, 정욱교. 2016. 위성 및 관측 자료를 이용한 지외선 지수 산출 연구. 한국기상학회지, 205-206.

가덕현, 오승택, 임재현. 2022. 태양객체 정보 및 태양광 특성을 이용하여 사용자 위치의 자외선 지수를 산출하는 DNN 모델. Journal of Internet Computing and Services(JICS). 23(2), 2287-1136(Online), 29-35.

김유근, 이화운, 문윤섭, 오하영. 2000. 파장별 자외선 예측모델을 이용한 복사속 추정. 한국기상학회지, 239-241.

이하나, 정욱교, 이원진, 이동원, 김준, 구자호. 2021. GEMS 관측자료를 이용한 동아시아 지역의 자외선 지수 및 보건지수 산출. 한국기상학회지, 250.

이하나, 김준, 정욱교. 2020. TROPOMI 자료를 이용한 동아시아 지역 자외선 지수 및 기타 보건지수 산출. 한국기상학회지, 215.

김미진, 김준, 윤종민. 2014. 천리안위성 기상 탑재체의 가시 채널 관측을 이용한 지표면 반사도 산출. Korean Journal of Remote Sensing Vol. 30, 5, 627-639.

김재정, 유용훈, 김창복. 2021. 기상 데이터와 기상 위성 영상을 이용한 다중 딥러닝 모델 기반 일사량 예측. J. Adv. Navig. Technol. 25(6), 569-575.

서현우, 이윤곤, 김창기. 2022. LSTM-GRU 기반의 일사량 예측 모델 성능 평가. 한국태양에너지학회지, 191.

기상청. 2020-2022. 자외선지수 자료.

기상청. 2020-2022. 기상위성 자료.

국가기상위성센터. "국가기상위성센터" 홈페이지. 2019. 구름 탐지 자료. https://nmsc.kma.go.kr/homepage/html/base/cmm/selectPage.do?page=static.edu.atbdGk2a

국가기상위성센터. "국가기상위성센터" 홈페이지. 2019. 지표면 반사도 자료. https://nmsc.kma.go.kr/homepage/html/base/cmm/selectPage.do?page=static.edu.atbdGk2a

국가기상위성센터. "국가기상위성센터" 홈페이지. 2019. 에어로졸 광학두께 및 에어로졸 입자 크기. https://nmsc.kma.go.kr/homepage/html/base/cmm/selectPage.do?page=static.edu.atbdGk2a