기상청

Weighted Extratree Ensemble 을 이용한
계절별 지면온도 예측 모델 개발

참가번호	230108	팀 명	

생활안전 (과제 1) – 기상에 따른 계절별 지면온도 산출기술 개발

1. 분석 배경 및 목표

지면온도는 시각과 계절, 기상이나 지표면의 상태에 따라 일사의 흡수, 장파복사, 대기와의 열교환 등이 달라 시공간적 차이가 크게 나타나는 기상요소이다. 지면온도는 국민의 실생활과 밀접한 관련이 있어 상세한 값이 필요하나, 지면온도 관측은 기온관측에 비해 훨씬 적은 지점에서 진행되고 있으며 직접 측정을 확대하기 위해서는 사회 · 경제적 비용 등이 발생한다는 한계점이 있다. 그러므로, 지면온도의 관측 공백을 해소하기 위해서는 다양한 기상자료를 활용한 계절별 지면온도 추정 기술 개발이 필요하다. 이에 본 공모안에서는 기상자료와 파생변수를 활용하여 계절별 지면온도 산출 모델을 개발하였다.

2. 분석 데이터 정의

본 공모안에서는 대회에서 제공한 데이터 중 관측소('stn')와 연도('year')를 제외한 변수를 사용하였다. 관측소와 연도는 학습 데이터와 테스트 데이터가 서로 구분되어 있어 제거하고 사용하였다. 범주형 변수(현천과 계절, 계절은 'mmddhh' 열을 가공하여 추가한 변수)는 원-핫 인코딩하였다.

계절별 모델 구축 과정에서 제공된 데이터를 가공하여 파생 변수를 생성했다. 가공하여 추가한 변수를 <표 1>에 노란색으로 표시하였으며, 그중 계절별로 선정한 변수는 <표 2>에 정리하였다. 데이터 전처리 및 변수 추가 과정에 대한 자세한 설명은 3 절에서 기술하였다.

변수명	정의	변수명	정의	변수명	정의
month	월	**date**	일	**hour**	시간
surface_tp_train.mmddhh	월, 일, 시	**surface_tp_train.ta**	1 시간 평균 기온(C)	**surface_tp_train.td**	1 시간 평균 이슬점온도(C)
surface_tp_train.hm	1 시간 평균 상대습도(%)	**surface_tp_train.ws**	1 시간 평균 풍속(m/s)	**surface_tp_train.rn**	1 시간 누적 강수량(mm)
surface_tp_train.re	1 시간 누적 강수유무(분)	**surface_tp_train.si**	1 시간 누적 일사량(MJ)	**surface_tp_train.ss**	1 시간 누적 일조량(초)
surface_tp_train.sn	위 시간: 00 분에 측정된 적설 깊이(cm)	**cos_hour**	시간변수 (코사인함수)	**sin_hour**	시간변수 (사인함수)
si.ano	계절별 si anomaly	**ss.ano**	계절별 ss anomaly	**ta.ano**	계절별 ta anomaly
si.avg	station 별 si anomaly 평균	**ss.avg**	station 별 ss anomaly 평균	**ta.avg**	station 별 ta anomaly 평균
si.year.avg	연도별 si anomaly 평균	**ss.year.avg**	연도별 ss anomaly 평균	**ta.year.avg**	연도별 ta anomaly 평균
index	전체 시간의 선형적 증가	**season**	계절	**surface_tp_train.ww**	현천 (C, F, G, H, R, S, X)

<표 1> 사용 변수 정의

계절	선정된 파생변수
봄	month, date, hour, cos_hour, sin_hour, si.avg, ss.avg, ta.avg, si.ano, ss.ano, ta.ano, index, season
여름	month, date, hour, cos_hour, sin_hour, si.avg, ss.avg, ta.avg, index, season
가을	month, date, hour, cos_hour, sin_hour, si.avg, ss.avg, ta.avg, si.ano, ss.ano, ta.ano, si.year.avg, ss.year.avg, ta.year.avg, season
겨울	month, date, hour, cos_hour, sin_hour, si.avg, ss.avg, ta.avg, si.year.avg, ss.year.avg, ta.year.avg, season

<표 2> 계절별 선정된 파생변수

3. 데이터 전처리 및 변수 추가

분석 데이터에 대한 그래프, 지면온도와의 상관계수 출력 등의 탐색적 자료분석(EDA) 과정을

바탕으로 결측치와 이상치를 처리하였으며, 제공된 데이터를 재가공한 변수를 추가하였다.

3.1. 결측치 처리

제공 데이터의 1 시간 평균 이슬점 온도(td), 상대습도(hm), 평균 기온(ta), 풍속(ws)에 존재하는 결측값 -99.9 를 NaN 으로 처리하였으며, 이후 이러한 결측값에 대하여 시간 선형 보간을 수행하였다. 위의 4 개 변수는 시간과 관련 있는 변수라고 판단하여 선형 보간을 수행하였으며, 나머지 변수는 결측값을 학습에 유의미하게 적용하는 트리 모델의 특성을 반영하여 결측값이 존재하여도 따로 처리하지 않았다. 한편, 종속 변수인 지면온도(ts)에 결측값이 있는 경우에는 해당 행을 제거하였다.

3.2. 계절을 구분하는 열 추가

계절별 지면온도 예측 모델의 학습을 위하여 데이터의 관측시점(월·일·시)을 나타내는 변수 'mmddhh'의 'mm'에 따라 2 ~ 4 월은 봄, 5 ~ 7 월은 여름, 8~10 월은 가을, 11~1 월은 겨울로 할당하는 새로운 변수 season 을 생성하여 기존 데이터에 추가하였다.

3.3. 시간 변수 추가

변수 'mmddhh'에서 월(month), 날짜(date), 시간(hour)에 대한 변수를 추출하였으며, 예측해야 하는 지면온도는 시간의 흐름에 따라 주기적 성질을 지니는 시계열 변수이므로 이를 반영하기 위해 주기성을 지닌 삼각함수 변수(sin_hour, cos_hour)를 추가하였다. 또한, 학습 및 검증데이터 전체에 대해 시간의 선형적 증가를 반영하고자 봄과 여름 모델에는 각 관측 지점에 대해 0 부터 1 씩 증가하는 index 열을 추가하였고, 가을과 겨울 모델에는 3.4 절에서 기술할 연도별 anomaly 평균열을 선택하여 추가하였다.

3.4. Anomaly 평균 변수 추가

검증 데이터에 관측 지점(stn) 및 연도(year)에 대한 정보가 주어지지 않았으므로, 이를 반영하여 적합하기 위해 학습 모델에 관련 변수를 포함해야 한다. 먼저 3.2 절의 과정으로 구분된 계절을 기준으로 학습 데이터를 분리한 후, 계절별로 일사량(si), 일조량(ss), 기온(ta)의 평균을 구하였다. 계산된 평균과 각 행의 원래 값과의 차이인 anomaly(si.ano, ss.ano, ta.ano) 열을 생성하여 봄과 가을 모델에 추가하였다. 이때 결측값 -99.9 는 평균 및 anomaly 계산 과정에서 제외하였다. 이후, 관측 지점의 위도 및 연도를 반영하기 위해 지점 · 연도별 anomaly 의 평균(지점 : si.avg,

ss.avg, ta.avg, 연도 : si.year.avg, ss.year.avg, ta.year.avg) 열을 추가하였다. 이때 지점별 anomaly 평균열은 모든 계절 모델에 추가하였으며, 연도별 anomaly 평균열은 가을과 겨울 모델에 추가하여 각각 계절별로 선형적인 시간의 흐름을 반영하였다.

3.5. 이상치 처리

예측해야 하는 종속 변수, 지면온도에 대해 box-plot 을 그린 후, 1.5 배 IQR 을 기준으로 상한과 하한을 구하였다. 지면온도 중 상한 값과 하한 값의 범위를 벗어나는 값들은 이상치로 간주하였으며, 이들을 1.1 배한 새로운 종속변수 ts2 를 생성하였다. 기존의 지면온도와 이상치에 가중치를 둔 지면온도를 각각 훈련 데이터로 할당하여, 이상 값에 대한 모델의 예측성을 높이고자 하였다.

4. 모델 구축

4.1. 엑스트라트리

엑스트라트리(Extratree)는 전체 훈련 세트를 사용하여 무작위로 결정 트리의 노드를 분할하는 트리 모델이다. 엑스트라트리는 무작위성이 커 여러 개의 결정트리를 훈련해야 하나, 노드 분할의 무작위성으로 인해 계산 속도가 높다는 특징이 있다. 본 공모안에서는 아래의 트리 앙상블 기반 모델 중 평가지표가 가장 우수한 엑스트라트리를 모델 구축에 사용하였다. 표는 3 절의 전처리 과정을 거친 Train 세트를 8:2 로 분할하여 각 단일모델로 지면온도 예측을 수행한 결과의 MAE 이다. MAE 가 낮은 모델은 AutoML 라이브러리인 pycaret 을 사용하여 5-fold 기준 상위 4 개를 선별하였다.

Model	평균 MAE(Mean Absolute Error)
Extratree	0.9961
Random Forest	1.1029
XGBoost	1.1752
LightGBM	1.3249

<표 3> 모델별 MAE

4.2. 앙상블 모델

하루 중 가장 높거나 낮은 지면온도에 대한 대한 예측성을 높인 엑스트라트리 앙상블 모델을 구축하였다. 엑스트라트리 단일모델로 예측을 수행한 결과, 모델이 하루 중 가장 높거나 낮은 지면온도를 과소모의함을 확인하였다. 따라서, 앞서 3.5 절에서 기술한 것과 같이 지면온도의 이상값을 극대화한 종속변수에 대해 학습한 엑스트라트리를 추가하여 앙상블 모델을 구축하였다. 앙상블 예측 결과는 식 (2)와 같이 두 개의 엑스트라트리 모델의 MAE 기반 가중치로 산출하였으며 그 산식은 아래에 기술하였다. 앙상블 모델 적용 결과, 전 계절에 걸쳐 이상값에 대한 예측성이 크게 개선됨을 확인하였다.

$$weight\ 1 = \frac{n}{\sum|(y1\ test)-(y1\ pred)|}\ ,\ \ weight\ 2 = \frac{n}{\sum|(y1\ test)-(y2\ pred)|} \quad \cdots (1)$$

$$weight\ pred = \frac{(weight1 \times y1\ pred + weight2 \times y2\ pred)}{(weight1 + weight2)} \quad \cdots (2)$$

	봄	여름	가을	겨울
Extratree 1	1.0651447257548787	0.9037065636346668	1.1945171206241025	1.4149194525803346
Extratree 2 (t2 종속변수 추가)	1.0367815179546294	0.8684021480659666	1.1287493044656363	1.388513244519171

<표 4> Extratree Ensemble Weight

앙상블 모델의 성능을 향상하기 위해 엑스트라트리의 하이퍼파라미터 튜닝을 진행하였다. GridSearchCV 를 사용한 경험적 하이퍼파라미터 튜닝 결과, 트리의 개수를 결정하는 파라미터 'n_estimators'는 가을 모델의 경우 500, 이외 계절(봄, 여름, 겨울) 모델의 경우 700 이 최적으로 나타났다. 더불어, 최상의 분할을 찾을 때 고려할 특징의 개수를 나타내는 파라미터인 'max_features'는 auto 로 설정하였다. 이외 파라미터는 Python 에서 제공하는 기본값을 사용하였다.

5. 모델 예측 결과

5.1. 최종 모델 선택

본 공모안에서는 엑스트라트리와 이상 값을 가중한 자료를 사용한 엑스트라트리의 앙상블

모델을 채택하여 최종 검증을 진행하였다. 계절별 최종 검증 결과와 평균 MAE 를 5.2 절에 기록하였다.

5.2. 최종 검증 결과

사용 모델	계절별 MAE (Mean Absolute Error)		평균 MAE
엑스트라트리 앙상블	봄	1.669	1.663
	여름	1.882	
	가을	1.443	
	겨울	1.658	

<표 5> 지면온도 예측 결과 MAE

6. 활용 방안 및 기대효과

지면온도는 국민의 실생활에 중요한 영향을 미치는 기상요소 중 하나이다. 따라서 시공간적으로 상세한 지면온도 수요가 증가하고 있으나, 관측소가 부족하다는 문제점이 존재한다. 이러한 상황에서, 기상자료를 활용한 계절별 지면온도 추정 기술은 각 지역에서 일상생활에 필요한 기상 정보를 생산하는 데 유용한 역할을 수행할 것이다. 특히 계절별로 기상 특성을 고려한 최적의 모델을 산출하였기 때문에, 여름철 일사량 예측이나 겨울철 결빙 예방 등 계절 특성에 따른 기상 정보 예측에 유용할 것으로 기대한다.

* 본 보고서 및 제출 코드(.ipynb)의 '봄'과 '여름' 모델은 윈도우 10 을 기반으로 한 기기에서 산출한 결과이고, '가을'과 '겨울' 모델은 Mac OS 를 기반으로 산출한 결과이다. 따라서 random state 가 OS 종류에 영향을 받으므로 다른 OS 에서 재검증 시 결과가 상이할 수 있다.

* 본 보고서와 함께 첨부한 파이썬 코드의 가을과 겨울 모델에서 코드 진행 편의 상 k-fold 결과로 산출한 가중치를 상수(float type)로 지정하였으나, 제출한 예측 값은 k-fold 로 산출한 원본(numpy.float64)으로 계산되어 타입 차이에 따른 미미한 값 차이가 있을 수 있다. 따라서 제출 예측 값에 대한 동일한 재현이 필요하다면 코드 별첨의 k-fold 실행이 필요하다.

참고문헌

구민호, 송윤호, & 이준학. (2006). 국내 지면 온도의 시공간적 변화 분석. 자원환경지질, *39*(3), 255–268.

박해선. (2020). 혼자 공부하는 머신러닝+딥러닝. 269.