기상청

머신러닝 알고리즘을 이용한 계절별 지면 온도 예측

참 가 번 호	230240	팀 명	zl존간zl

김자현*· 김재훈**· 나상우***·백대환****

*고려대학교(세종) 공공정책대학 빅데이터전공

**고려대학교(세종) 공공정책대학 빅데이터전공

***고려대학교(세종) 공공정책대학 빅데이터전공

****고려대학교(세종) 공공정책대학 빅데이터전공

1. 서론

최근 지구온난화로 인해 폭염이나 한파 등의 전 지구적 이상기후 현상이 빈번히 발생한다. 이로 인해 작물 생산에 어려움이 있다. 특히 토양의 온도는 식물의 생육, 미생물의 활동, 토양생성작용 등에 중요한 요소이다. 토양온도가 낮아지면 유기물의 분해가 늦어져 다량으로 쌓이게 되고, 온도가 높아지면 유기물의 분해가 빨라 무기화작용이 촉진된다. 그러므로, 식물의 성장에는 지표 온도를 관측하여 조치를 치하는 것이 중요하다. 기존의 지면온도 측정 방법에서는 온도계 구부가 지면에 노출되지 않을 정도로 낮게 묻고 모세관 윗부분이 약간 위로 치켜지도록 지주를 세워 둔다. 하지만 날씨가 좋지 않을 때 바람과 비로 인하여 온도계의 수감 부가 노출되기 쉽고, 일사나 지면 상태에 따른 변화가 심하여 일정한 상태를 유지하면서 관측한다는 것은 매우 어렵다. 따라서 인공위성 데이터를 이용하여 지면온도를 예측하면 기상 날씨에 영향을 받지 않고 기록할 수 있다. 지면온도의 관측 값은 농업기상의 한 요소로 중요할 뿐만 아니라 토목, 건축 등 각 산업분야에 널리 이용되길 기대한다.

2. 활용데이터

2.1 종류 기상청(과제 1 제공 데이터)

2.2 범위

① 학습 데이터 A년 2월~E년 1월(5년 동안의 기상데이터 + 지면온도)

② 검증 데이터 F년 2월~G년 1월(1년 동안의 특정3지점의 기상데이터)

2.3 기상 관측데이터 항목별 구성

데이터의 칼럼들 이름과 설명은 <**Table 1**>와 같다.

종속변수는 빨간색으로 표시하였다.

<Table 1> 데이터 칼럼 이름 및 설명

YYYY	년도
MMDDHH	월/일/시간
STN	지점번호
TA	1시간 평균기온(C)
TD	1시간 평균 이슬점온도(C)
HM	1시간 평균 상대습도(%)
WS	1시간평균풍속(m/s)
SI	1시간 누적 일사량(MJ)
SS	1시간 누적 일조량
RN	1시간 누적 강수량
SN	위 시간:00분에 측정된 적설 깊이(cm)
RE	1시간 누적 강수유무(분)
WW	현천계 현천(S:눈, R:비, F:안개, H:박무, G:연무, C:맑음, X:모름)
TS	**1시간 평균 지면온도(C)**

3. 데이터 탐색 및 데이터 전처리

3.1 상관분석

종속변수 TS(지면온도)와 다른 칼럼과의 상관관계는 <**Table 2**>와 같다.

<Table 2> 지면온도와 다른 칼럼들에 대한 상관계수

컬럼	TA	TD	SI	SS	WS	MMDDHH	SN	HM	STN	RE	RN
지면온도	0.92	0.78	0.36	0.36	0.25	0.16	0.16	0.15	0.09	0.09	0.06

3.2 결측치 분석

-99.9, -99인 값을 Numpy 패키지를 이용하여 np.nan 값으로 대체 후 비율을 확인하였다.

전체 데이터에 대한 결측치 비율은 <**Table 3**>와 같다.

<Table 3> 각 칼럼들에 대한 전체 결측치 비율

컬럼	TA	TD	SI	SS	WS	MMDDHH	SN	HM	STN	RE	RN
지면온도	0.14	0.15	45.6	45.4	0.23	0.00	98.4	0.13	0.00	0.49	1.69

이후, RN(1시간 누적 강수량), RE(1시간 누적 강수 유무), SI(1시간 누적 일사량), SS(1시간 누적 일조량), SN(00분에 측정된 적설 깊이) 위 6가지의 결측치는 0으로 대체하였다.

TA(1시간 평균 기온), TD(1시간 이슬점 온도), HM(1시간 평균 상대습도), WS(1시간 평균 풍속), TS(1시간 평균 지면온도) 위 5가지의 결측치는 Imputation 기법 중 Spline interpolation(스플라인 보간법)을 이용하여 대체하였다.

Spline Interpolation: 결측치 주변의 데이터 패턴을 파악하여 부드러운 곡선을 형성하는 기법이다. 결측치 위치에서의 예측값을 생성하기 위해 주변 데이터의 패턴을 활용하고, 구간별로 다항식을 이용하여 데이터를 부드럽게 연결하는 방식으로 구성된다.

4. 모델 선정 과정

여러가지 모델을 사용하여 각 계절별로 최적의 모델을 찾아보았다.

① **MultipleLinearRegression**

종속변수와 여러 개의 독립변수 간의 선형 관계를 모델링하는 회귀분석 기법이다.

② **RandomForestRegression**

의사결정트리를 기반으로 하는 앙상블 기법이다. 여러 개의 의사결정트리를 생성하고 각각의 트리에서 예측한 결과를 최종 예측값으로 얻는 기법이다.

③ **GradientBoostRegression**

여러 개의 결정트리를 순차적으로 학습시켜 오차를 보정해나가는 앙상블 기법이다. 각 트리는 이전 트리의 오차를 줄이기 위해 학습되며 강력한 예측 모델이 되는 기법이다.

④ **XGBoostRegression**

Gradient Boosting 알고리즘을 기반으로 한 강력하고 효율적인 앙상블 기법이다.
트리기반이고, 예측 성능과 실행 속도를 모두 고려한 최적화가 이루어져 있다.

⑤ **CatBoostRegression**

범주형 변수를 자동으로 처리할 수 있는 부스팅 알고리즘 모델의 일종으로, 범주형 변수의 처리를 최적화하여 모델의 예측 성능을 향상시킨다.

위 5가지 모델 중 가장 성능이 좋은 CatBoostRgression 모델을 사용해 하이퍼파라미터 최적화를 진행하였다. CatBoostRgression의 특징은 다음과 같다.

자동 범주형 변수 처리:

CatBoost는 범주형 변수를 자동으로 처리할 수 있어, 데이터 전처리 과정에서 범주형 변수를 수치형으로 변환하는 번거로움을 줄일 수 있다. 이는 특히 범주형 변수가 많은 데이터 셋에 대한 처리를 단순화하고, 실수를 줄여주는 장점을 가지고 있다.

자동 특성 스케일링:

CatBoost는 모델 학습 시 자동으로 특성 스케일링을 수행한다. 이는 각 피쳐의 스케일 차이로 인한 모델 성능 저하를 방지하고, 예측 모델의 안정성과 성능을 향상시키는 데 도움을 준다.

고성능 부스팅:

CatBoost는 그래디언트 부스팅 알고리즘을 기반으로 하여, 속도와 성능 면에서 우수한 결과를 제공한다. 다양한 손실 함수와 평가 지표를 지원하여 다양한 문제에 적용 가능하다.

과적합 방지:

CatBoost는 과적합을 방지하기 위한 다양한 기법을 사용한다. 또한, 하이퍼파라미터를 통해 모델의 일반화 성능을 향상시킬 수 있다. 이는 훈련 데이터에 과도하게 적합하지 않으면서도 일반적인 패턴을 잘 캡처할 수 있도록 돕는다.

이러한 기능들은 CatBoost가 다양한 문제와 데이터 세트에 유연하게 적용될 수 있게 하며, 이 연구에서도 이 알고리즘의 성능을 평가하고 활용하였다.

5. 지면온도 예측 모델 구성

5.1 데이터 구성

5.1.1 사계절별 데이터 구성

봄 하이퍼파라미터 데이터, 훈련용 데이터와 테스트 데이터는 **<Table 4>**와 같이 구성된다.

<Table 4> 봄 훈련용, 테스트 데이터셋 구성

DATA TYPE	STN	YEAR	DATA
Hyper-Parameter	10	2월 ~ 4월, 5년	10740
Train	10	1월 ~ 12월, 5년	438240
Test	3	2월 ~ 4월, 1년	6405

봄의 경우, 특수하게 훈련용 데이터로 하이퍼파라미터 설정에 사용된 데이터와 다르게 1월부터 12월 데이터 전체를 사용하였다.

여름 훈련용 데이터와 테스트 데이터는 **<Table 5>**와 같이 구성된다.

<Table 5> 여름 훈련용, 테스트 데이터셋 구성

DATA TYPE	STN	YEAR	DATA
Train	10	1월 ~ 12월, 5년	110400
Test	3	2월 ~ 4월, 1년	6619

가을 훈련용 데이터와 테스트 데이터는 **<Table 6>**와 같이 구성된다.

<Table 6> 가을 훈련용, 테스트 데이터셋 구성

DATA TYPE	STN	YEAR	DATA
Train	10	8월 ~ 10월, 5년	110400
Test	3	8월 ~ 10월, 1년	6624

겨울 훈련용 데이터와 테스트 데이터는 **<Table 7>**와 같이 구성된다.

<Table 7> 겨울 훈련용, 테스트 데이터셋 구성

DATA TYPE	STN	YEAR	DATA
Train	10	11월 ~ 1월, 5년	110400
Test	3	11월 ~ 1월, 1년	6624

각 계절에 특성에 따라 다른 데이터를 사용하여 모델의 하이퍼파라미터를 최적화하는 과정을 수행하였다. 각 계절마다 다른 기후 패턴과 특성을 가지므로, 각 계절에 최적화된 모델을 만드는 것이 중요하다.

5.2 CatBoost 하이퍼파라미터 최적화

5.2.1 봄

봄 계절에 대한 모델 학습에서는 특별한 접근법을 사용하였다. 봄의 특성에 맞게 하이퍼파라미터를 최적화한 후, 전체 데이터를 이용하여 모델을 학습시켰다. 이는 봄 계절의 기후 특성과 데이터 분포를 반영하여 모델의 일반화 성능을 향상시키기 위함이다.

5.2.2 여름, 가을, 겨울

나머지 계절인 여름, 가을, 겨울에 대해서는, 각 계절의 특성에 맞게 데이터를 선택하고 이를 바탕으로 하이퍼파라미터를 최적화하였다. 이후, 선택된 훈련 데이터만을 이용하여 모델을 학습시켰다. 이는 각 계절별로 다른 특성을 가진 데이터를 통해 더 정확한 예측 성능을 도출하기 위한 전략이다.

이렇게 각 계절에 맞는 데이터를 바탕으로 하이퍼파라미터를 최적화하고 모델을 학습시키는 방법은 각 계절별로 다른 기후 특성과 패턴을 잘 반영할 수 있게 해주며, 이를 통해 모델의 예측 성능을 최적화할 수 있다. 이 방법은 계절별 기후 예측의 정확도를 높이는 데 큰 도움을 주었다.

5.3 하이퍼파라미터 설정

초기 모델 훈련 과정에서 성능을 최적화하기 위해 다양한 하이퍼파라미터 설정이 필요하였다. 이를 위해 하이퍼파라미터 최적화를 위한 오픈 소스 프레임워크인 Optuna를 사용하였다.

Optuna를 사용하여 CatBoostRegressor 모델의 각종 하이퍼파라미터를 조정하였다. 주요하게 조정한 하이퍼파라미터들은 다음과 같다.

- 'iterations' : 최적화를 위한 반복 횟수를 설정하였다. 이는 100에서 15000 사이의 값을 가지도록 설정하였다.
- 'od_wait' : 과적합 탐지를 위한 대기 시간으로, 500에서 2300 사이의 값을 설정하였다.
- 'learning_rate' : 모델 학습률을 결정하는 하이퍼파라미터로, 0.01에서 1 사이의 값을 설정하였다.
- 'reg_lambda' : L2 정규화를 위한 하이퍼파라미터로, 1e-5에서 100 사이의 값을 설정하였다.
- 'subsample' : 데이터의 부분 샘플링을 결정하는 비율로, 0에서 1 사이의 값을 설정하였다.
- 'random_strength' : 임의의 강도를 결정하는 하이퍼파라미터로, 10에서 50 사이의 값을 설정하였다.
- 'depth' : tree의 깊이를 결정하는 하이퍼파라미터로, 1에서 15 사이의 값을 설정하였다.
- 'min_data_in_leaf' : 트리의 리프 노드에서 필요한 최소 데이터 개수를 설정하는 하이퍼파라미터로, 1에서 30 사이의 값을 설정하였다.
- 'leaf_estimation_iterations' : 리프 추정을 위한 반복 횟수를 설정하는 하이퍼파라미터로, 1에서 15 사이의 값을 설정하였다.
- 'bagging_temperature' : 배깅의 temperature를 결정하는 하이퍼파라미터로, 0.01에서 100.00 사이의 로그-균등 분포 값을 설정하였다.
- 'colsample_bylevel' : tree의 level 별로 feature를 샘플링하는 비율을 설정하는 하이퍼파라미터로, 0.4에서 1.0 사이의 값을 설정하였다.

이러한 하이퍼파라미터들을 설정한 후, 모델을 훈련 데이터에 적합시키고 평가를 위해 테스트 데이터로 Mean Squared Error(MSE)를 계산하였다.

Optuna의 study를 생성하고, 최적의 하이퍼파라미터를 찾기 위해 총 100회의 시도를 하였다. 이 과정에서 Optuna는 지정된 방향('minimize')에 따라 하이퍼파라미터를 조정하며 MSE를 최소화하는 방향으로 최적화를 수행하였다. 최종적으로 모델의 성능을 최적화하는 데 가장 효과적인 하이퍼파라미터 조합을 찾을 수 있었다.

사계절 하이퍼파라미터는 **<Table 8>**과 같이 설정된다.

<**Table 8**> 사계절 하이퍼파라미터 설정

Parameter	**봄**	**여름**	**가을**	**겨울**
iterations	4417	2102	112	11836
od_wait	901	1998	1004	618
learning_rate	0.019	0.01	0.13	0.01

reg_lambda	34.1	58.68	51.78	45.24
subsample	0.11	0.32	0.14	0.25
random_strength	16.91	13.14	17.61	16.63
depth	12	5	8	5
min_data_in_leaf	5	17	7	8
leaf_estimation_iterations	8	3	4	15
bagging_temperature	0.19	0.02	0.23	0.23
colsample_bylevel	0.76	0.90	0.74	0.66

5.4 검증 결과

모델의 일반화 능력을 평가하고, 실제 환경에서의 성능을 추정하는 단계다.

<Table 9> 에 나타난 검증 결과를 통해 계절별로 모델의 성능이 어떻게 변화하는지 살펴보았다. 봄, 가을에 대한 예측에서는 모델이 우수한 1.665, 1.669의 Mean Absolute Error(MAE) 점수를 기록하였다. 이는 모델이 봄 시즌의 데이터에 대해 매우 정확한 예측을 수행하였음을 나타낸다. 반면에, 여름, 겨울에 대한 예측에서는 2.015, 2.151로 다소 낮은 성능을 보였다. 이는 여름, 겨울 데이터의 패턴이 복잡하거나, 학습 데이터가 충분하지 않아 모델이 데이터에 대한 예측을 덜 정확하게 수행했음을 시사한다. 이를 통해 모델의 성능은 계절에 따라 다르며, 특히 여름 데이터에 대한 예측 성능 향상이 필요함을 확인하였다.

<**Table 9**> 사계절 MAE 검증 결과

Season	MAE Score
봄	1.665
여름	2.015
가을	1.669
겨울	2.151
평균	1.875

6. 결론

데이터 탐색 및 분석 과정에서 초기에 발견된 기상청 데이터의 문제점들은 효율적인 전처리를 통해 보완되었으며, 이는 모델 성능 향상에 중요한 역할을 하였다. 이후, CatBoost 알고리즘을 이용하여 예측 모델을 설계하였으며, 이 결과 사계절 평균 1.875의 MAE 값을 얻었다. 여름과 겨울에 대한 예측 성능은 봄과 가을에 비해 MAE 값이 약 0.35~0.5 정도 낮은 성능을 보였다. 파라미터를 조정함으로써 성능을 일정부분 향상시킬 수 있지만, 이에는 한계가 존재함을 인지하였다.

본 연구 결과는 기상 예측의 정확도를 향상시키는 데 기여하며, 이로 인해 다양한 산업 분야에 긍정적인 영향을 끼칠 수 있음을 보여준다. 그러나, 아직 예측 성능의 개선 여지가 존재하므로, 추가적인 연구가 필요함을 시사한다.