기상청

기상에 따른 계절별 지면 온도 산출기술 개발

[아침기상] 김수빈 김준서 박시언 서희나 이지윤 조웅빈

TABLE OF CONTENTS

공모배경

지면온도란?

지표면 부근에서 측정한 온도로
시각과 계절에 따른 차이 뿐만 아니라
기상이나 지표면의 상태에 따라 일사의 흡수, 장파복사,
대기와의 열 교환 등이 현저히 다르므로
시공간적인 차이가 크게 나타날 수 있는 기상요소

노지(밭) 자동관측장비 | 노지(밭) 지면온도 센서 (붉은색 원)

지면온도 측정의 중요 사례

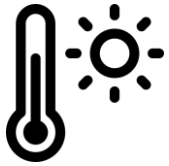
여름철 기온은 34.7도 임에도 불구하고 노지의 지면온도는 51.3도까지 오름

아침 기온은 영상권을 회복했음에도, 지면 온도가 영하권에 머물면서 도로가 살얼음판이 됨

폭염 경보 및 결빙 알림은 기온 뿐만 아니라 **지면온도에 대한 고려가 필요**함

분석 목표

기상에 따른 계절별 지면온도 산출기술 개발의 필요성



지면온도 예측은 국민
실생활과 밀접한 관련 존재

시공간적으로 상세한 지면온도
예측 수요는 날마다 증가

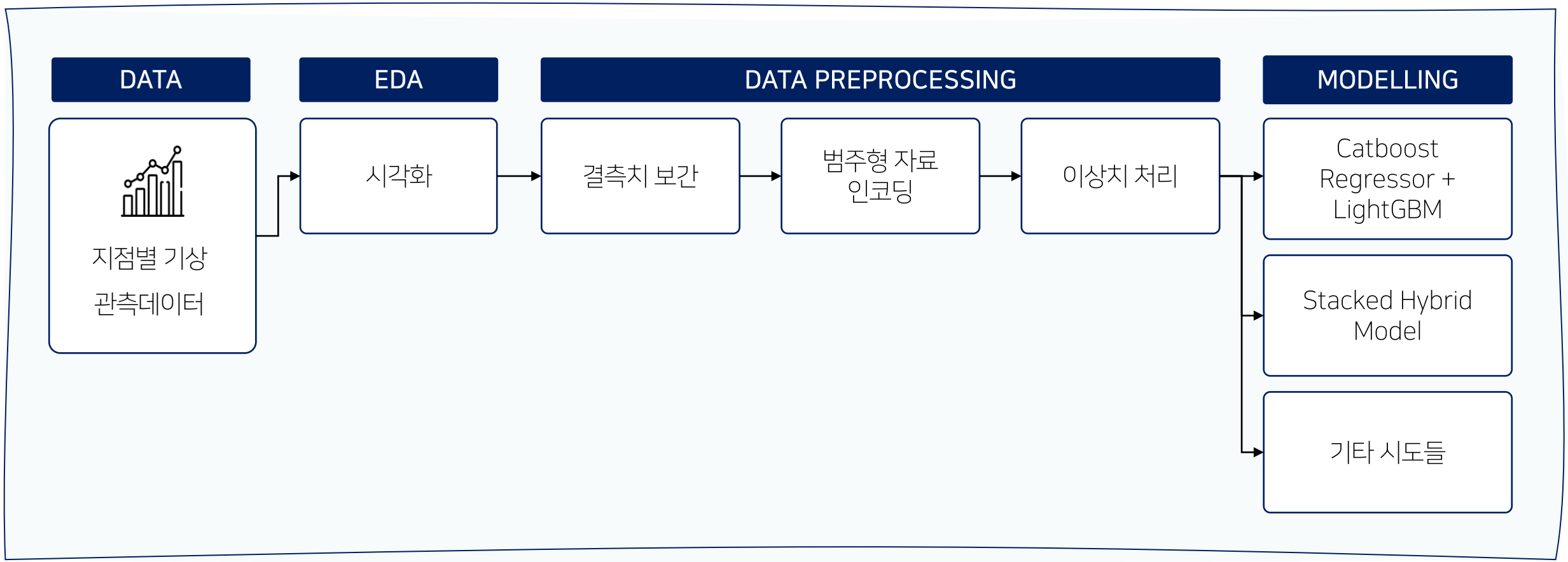
기온 관측 지점에 비해
지면온도는 **훨씬 적은 지점에서 관측**하고 있기 때문에
기상 자료를 활용한 **지면온도 추정 기술이 필요**

분석 목표

① 주어진 10개 지점에 대한 11개의 기상 관측 데이터를 활용하여
새로운 지점의 지면 온도를 예측하는 **계절별 지면온도 총합 산출 모델** 개발

② 예측 모형을 활용하여 실현 가능한 활용방안을 제시

분석 흐름



해당 과정을 통해 계절별 지면온도 총합 산출 모델을 개발하고자 함

데이터 정의

테이블명	내용	테이블명	내용
YYYY	년도	HM	1시간 평균 상대습도
MMDDHH	월/일/시간	HM	1시간 평균풍속
STN	지점번호	SI	1시간 누적 일사량
TA	1시간 평균 기온	SS	1시간 누적 일조량
TD	1시간 평균 이슬점 온도	RN	1시간 누적 강수량
WW	현 천계 현천	RE	1시간 누적 강수유무(분)
SN	적설 깊이	TS	1시간 평균 지면 온도

고려 항목

① -99, 99.9 , -999와 같은 결측치가 다수 존재하여
추후 결측치 처리를 위해 NA값으로 변경

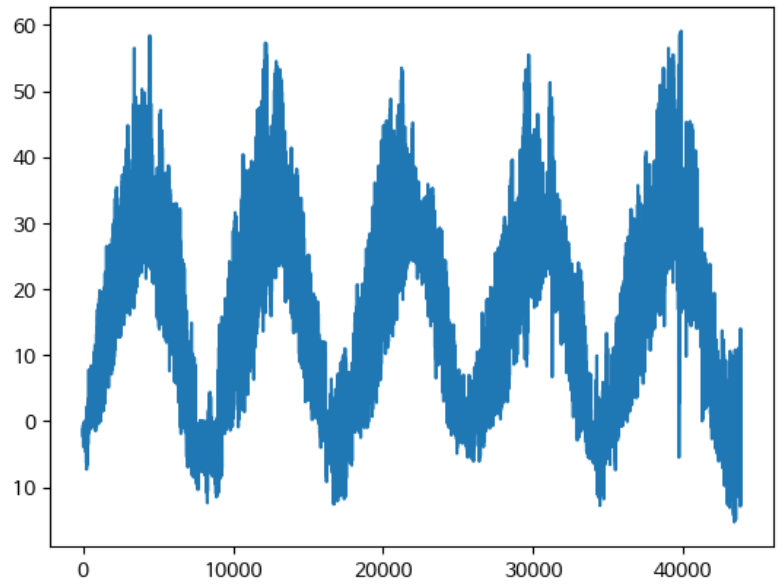
② "새로운 관측지점에 관한 계절별 지면 온도 예측"
→ 분석 과제를 고려한 다양한 교차 검증 방식 채택

10개의 기상 관측 지점에 대한 다년간의 기상 관측 데이터로 **지점별 기상 관측데이터 항목별 구성**
지점번호, 년도, 월/일/시간 등 기상 관측 지점과 시점에 대한 정보는 비공개

탐색적 자료 분석 (EDA)

: 시계열 특성 시각화 (TS Plot)

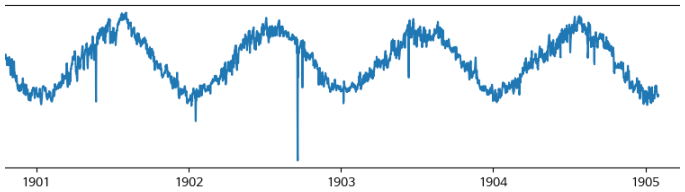
지점의 지면온도 시계열 플랏



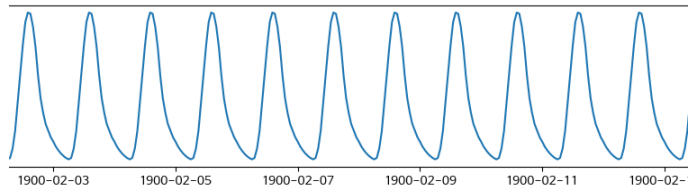
모든 10개 지점의 지면온도에 대해
추세 및 계절성이 존재하는 **비정상시계열**임을 확인

시계열적 요소 분해

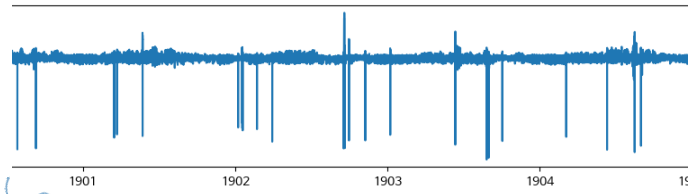
[Trend] 경미한 1차 추세 존재



[Seasonality] 강한 계절성



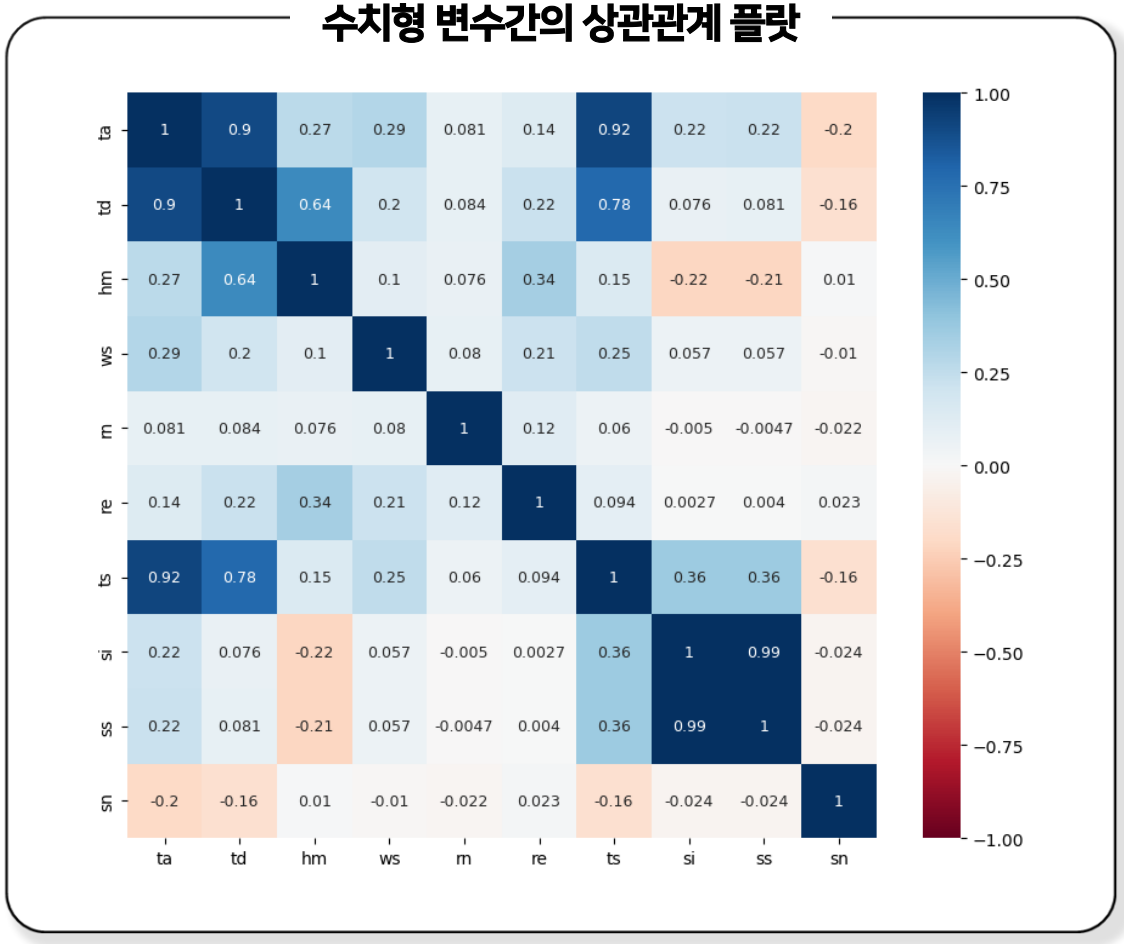
[Residual] 랜덤한 분포



순환요인(cycle) 존재 X

탐색적 자료 분석 (EDA)

: 수치형 변수 간 상관관계



High Correlation

평균 기온(TA), 평균 이슬점온도(TD), 평균지면온도(TS)
누적 일사량(SI), 누적 일조량(SS)

① 지면온도(TS)가 높을수록, 평균 기온(TA)와
평균이슬점온도(TD)도 높아짐

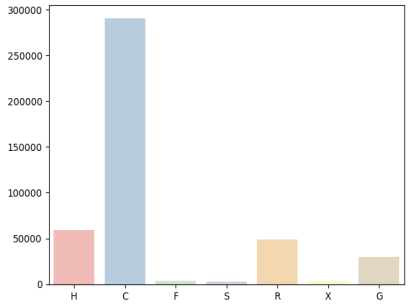
② 누적 일사량(SI), 누적 일조량(SS)
두 변수의 경향성이 매우 유사

탐색적 자료 분석 (EDA)

: 범주형 / 수치형 변수 시각화

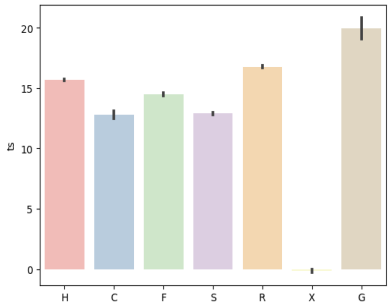
범주형 변수

현천계 현천(WW) 유형별 개수



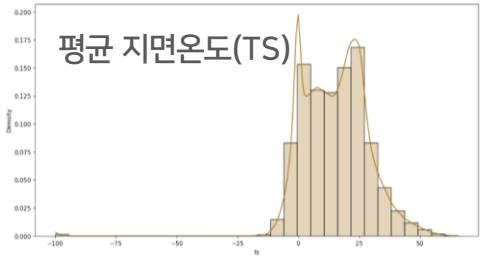
맑은(C) 날이 가장 많음

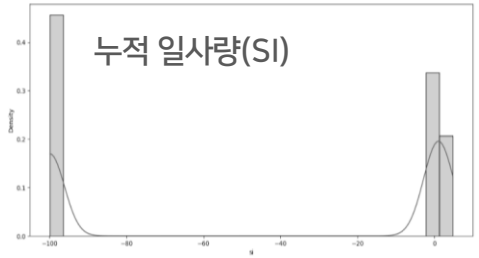
WW 유형별 평균 지면온도(TS)



연무(G) 평균 지면온도가
가장 높지만, 신뢰도가 떨어짐

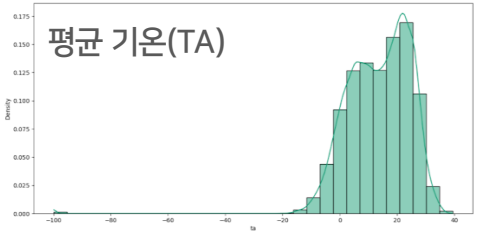
수치형 변수 A

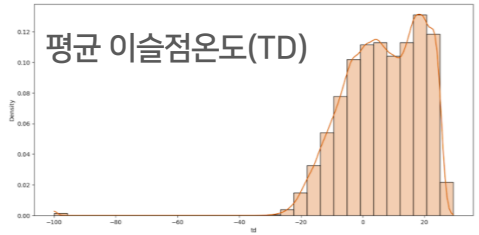


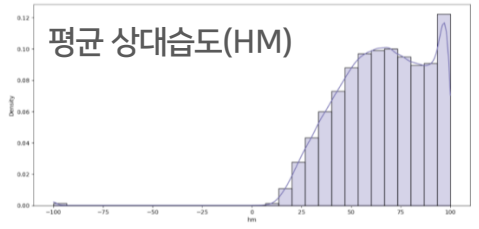


수치형 변수 B

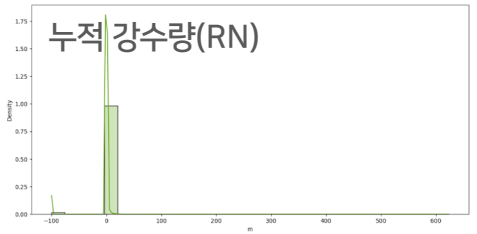
유사한 분포 형태

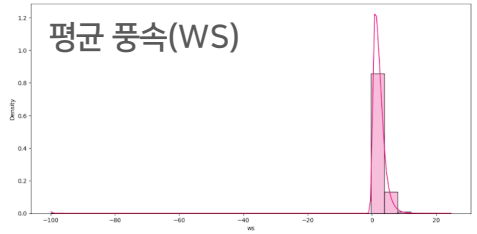


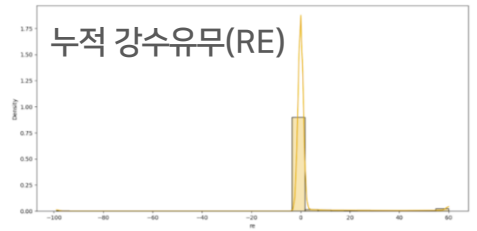




치우친 분포





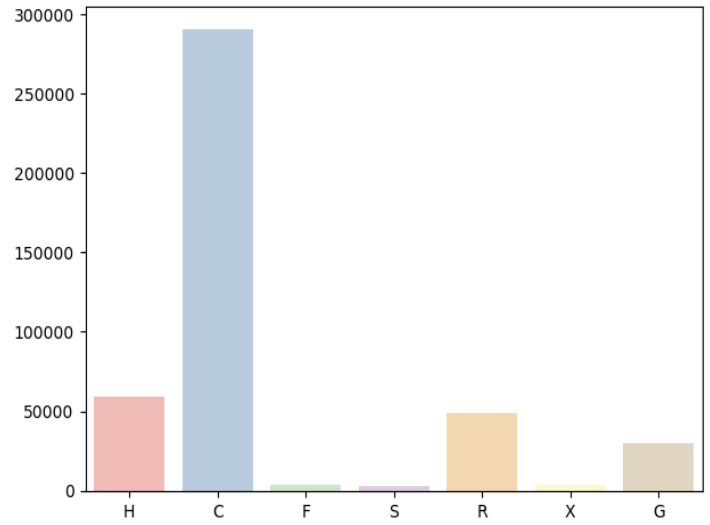


항목 시각화와 분포 시각화를 통해 범주형, 수치형 변수들의 경향성을 확인

탐색적 자료 분석 (EDA)

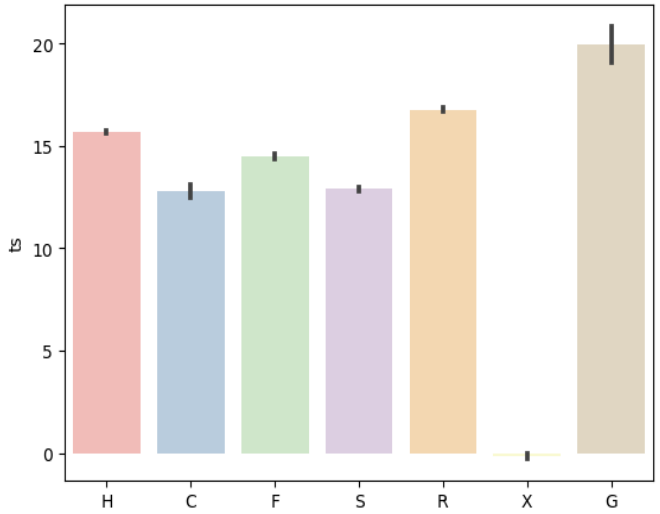
범주형 변수 시각화

현천계 현천(WW) 유형별 개수



맑은(C) 날이 가장 많음

WW 유형별 평균 지면온도(TS)



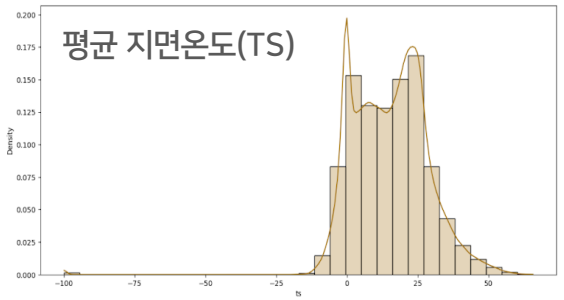
연무(G) 평균 지면온도가 가장 높지만, 신뢰도가 떨어짐

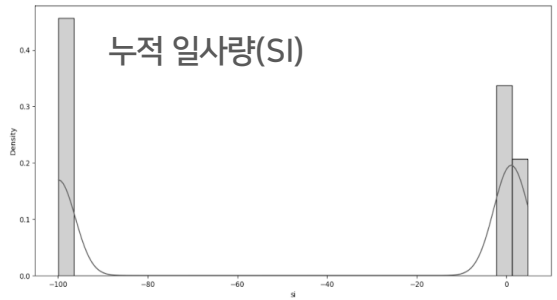
맑은 날과 연무 일 때 지면온도가 가장 높지만 신뢰성은 없음

탐색적 자료 분석 (EDA)

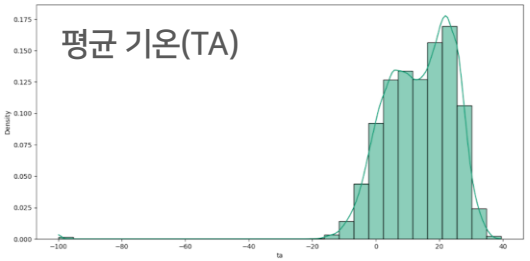
수치형 변수 시각화

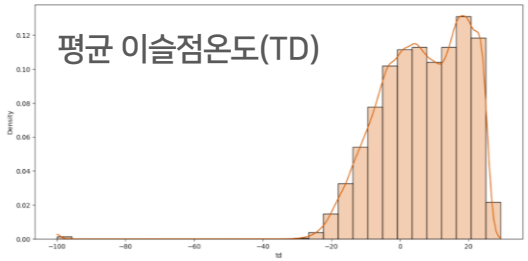
Target & SI

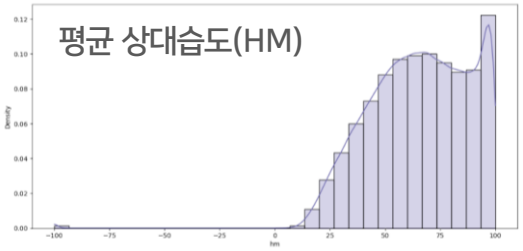




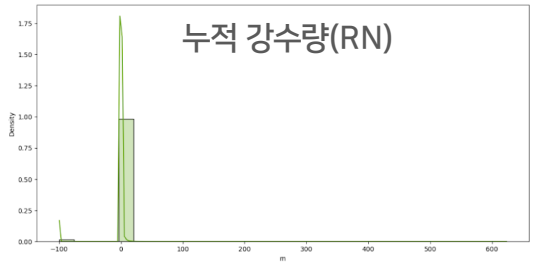
유사한 분포의 형태

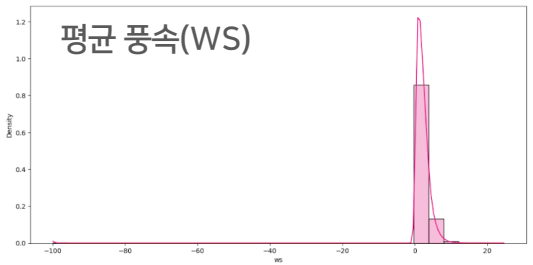


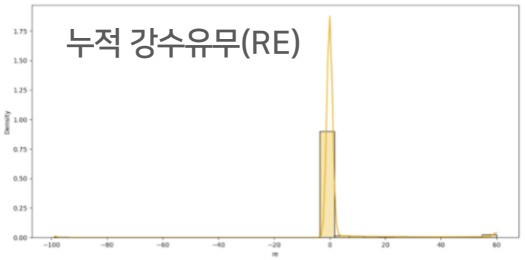




치우친 분포의 형태

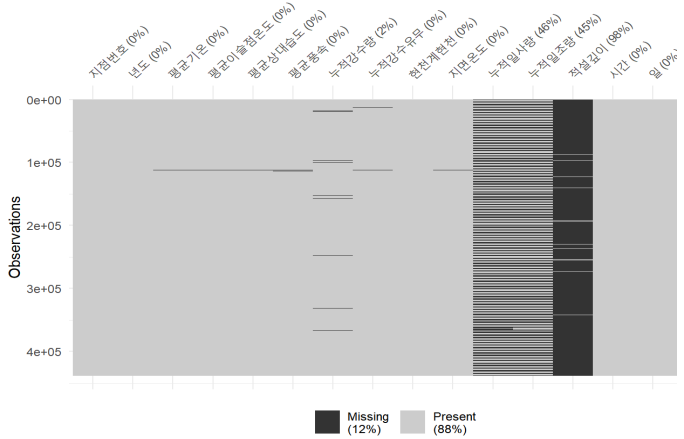




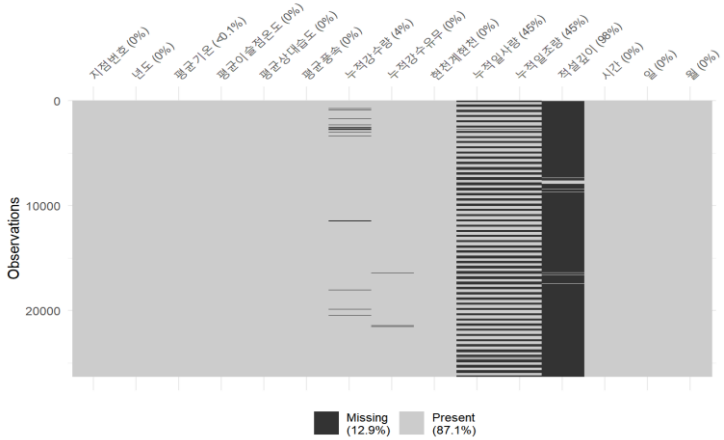


수치형 변수들의 분포는 크게 ① 유사한 분포 ② 치우친 분포의 형태로 나타남

결측값(Missing Value) 처리



학습데이터의 결측치 시각화



검증데이터의 결측치 시각화

결측치 비율 파악

종속변수를 포함한

9개의 수치형 변수에서 결측 발생

↓

누적 일사량(SI), 누적 일조량(SS), 적설 깊이(SN)의

결측치 비율 40% ↑

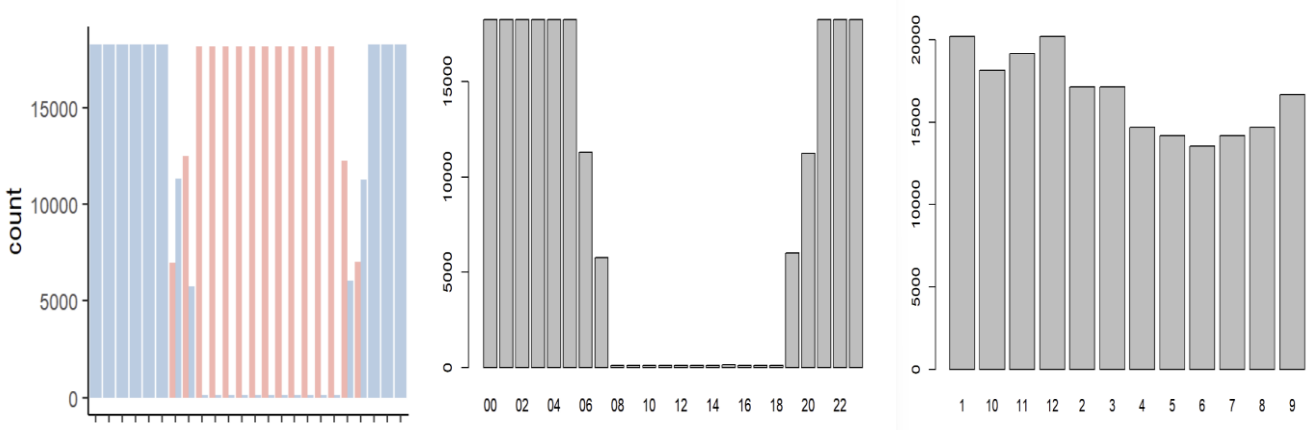
결측치 발생 패턴 파악

결측치에 발생 패턴이 존재하는지 확인하기 위해 ① 결측치인 경우 ② 결측치가 아닌 경우로 나누어 시각화 진행

→ 결측치 **비율이 낮은 변수의 경우 결측치 발생 패턴 뚜렷하지 않음**

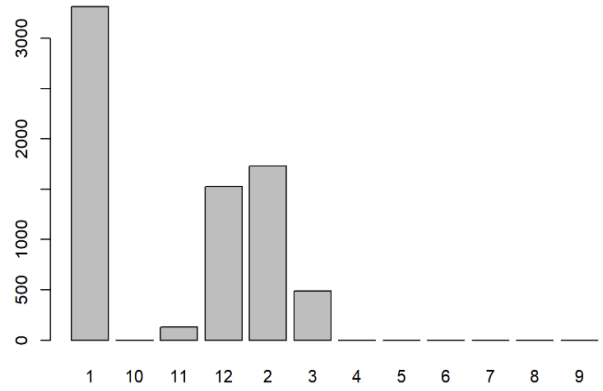
결측값(Missing Value) 처리

시간/월에 따른 1시간 누적 일사량/일조량 결측치



누적 일사량(SI)와 일조량(SS) → 밤 시간대에 주로 결측값이 발생한 것을 확인

월별 적설 깊이 결측치

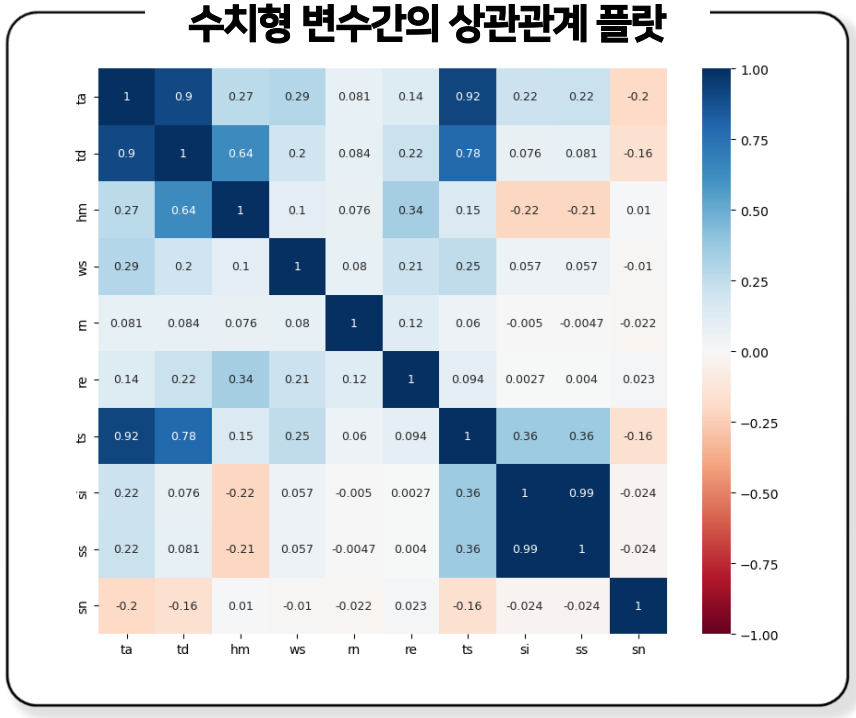


눈이 오지 않는 계절의 경우 → 결측치 발생

결측치 비율이 높은 변수의 경우, 결측값의 발생 패턴이 존재함을 확인 (**MNAR** , Missing Not At Random)

따라서 SI ,SS , SN의 결측치는 학습 데이터와 검증데이터 모두 **0으로 대체**

결측값(Missing Value) 처리



Bayesian Ridge

Scikit-learn에서 진행한 Imputer 성능 비교 연구에서
Iterative Imputation + Bayesian Ridge 보간이
가장 원본 데이터와 비슷하게 결측치를 보간한 결과 존재

데이터 왜곡 방지

① Data Leakage 방지를 위해 Iterative Imputer 적합 시 학습데이터의 정보만 이용

② 종속변수인 TS의 결측값을 보간할 경우 데이터의 왜곡이 발생할 것이라 판단

→ TS에서 결측값이 발생한 행을 삭제

나머지 변수의 경우 상관관계가 높은 변수 조합을 고려해 Scikit-learn의 Iterative Imputer의
Bayesian Ridge를 사용해 결측치 보간

이상값 및 범주형 변수

이상값(Outlier) 처리

예측 모델링의 평가지표인 MAE는 이상치에 강건함
지면온도 추정에 있어, 데이터의 이상치 역시 의미 있을 것이라 판단
→ **이상치의 영향을 최소화**하는 방법 고려

Robust Scaling

데이터의 중앙값이 0, IQR(Inter Quantile Range)=1이
되도록 스케일링하는 기법

현천계 현천(WW) Encoding

현천계 현천(WW)은 7개의 범주를 수치형 변수로 전환

S: 눈, R: 비, F: 안개, H: 박무, G: 연무, C: 맑음, X: 모름

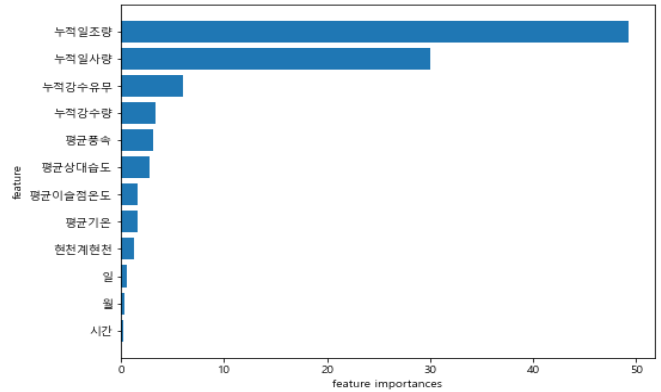
Label Encoding/One-Hot Encoding

Label Encoding: 코드형 숫자값으로 변환
One-Hot Encoding: 고유값 칼럼은 1, 나머지 칼럼은 0으로 표시

이상치의 영향을 최소화하는 Robust Scaling 진행 & 모델링 방법에 따라 현천계 (WW) Encoding 진행

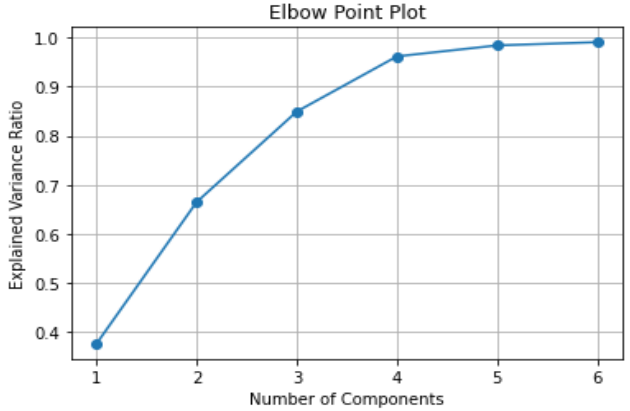
변수선택 및 차원축소

변수 선택



분석 목적에 부합하는 소수의 예측 변수만을 선택

PCA



데이터의 분산 구조를 잘 설명하는

축을 구하는 방식으로 차원축소를 진행

변수 선택 시 모델의 성능이 급격하게 떨어져 채택하지 않음

주성분분석 전 시각화한 Scree Plot의 형태가 일직선으로 나타남 → 차원축소가 효과적이지 않을 것이라 판단

다양한 시도

Machine Learning

Linear Model

Lasso, Ridge, Bayesian Ridge,
Elastic-Net, MLP Regressor

오버피팅 및 이상치에 영향 많이 받음

Boosting Model

XGBoost, LGBM, CatBoost,
HistGradientBoosting Regressor

선형 모델보다는 성능이 좋지만,
Peak 및 이상치를 여전히 추정을 잘 못함

Time Series and Other

SARIMAX: 모델 적합이 느림

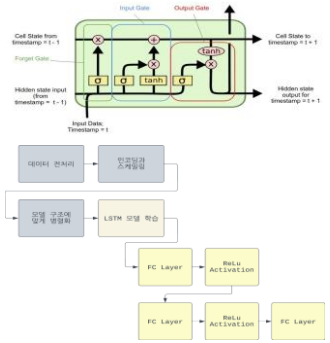
Pycaret AutoML: 모델 적합이 느리며,
단일 모델에 비해 성능 향상 X

NA값 처리, Robust Scaling, Label Encoding을 한 경우 및 안 한 경우에 대해 모두 모델링을 진행했으며
이때, CV는 상황에 따라 Time Series CV, K-Fold 등을 사용함

다양한 시도

Deep Learning

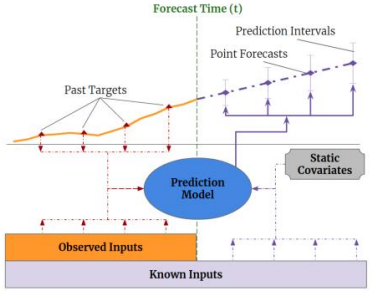
LSTM

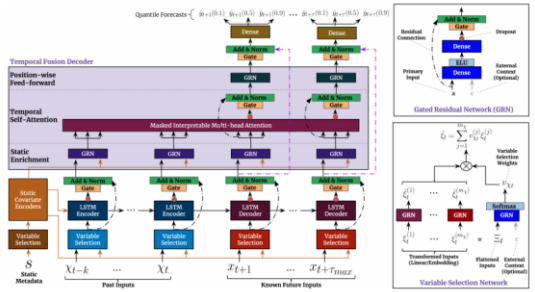


지점	시점	…	지면온도
1	A-01-01		a
2	A-01-02		b
…	…		…
10	C-12-31		c

한번에 삽입할 때, 10개의 지점, 7개의 변수들이 한번에 병렬
→ 모델 내 지점 앙상블 효과
LSTM을 사용하여 sequential 특징 유지
Validation MAE는 1.8정도 유지, Test MAE에서는 BAD

Temporal Fusion Transformer



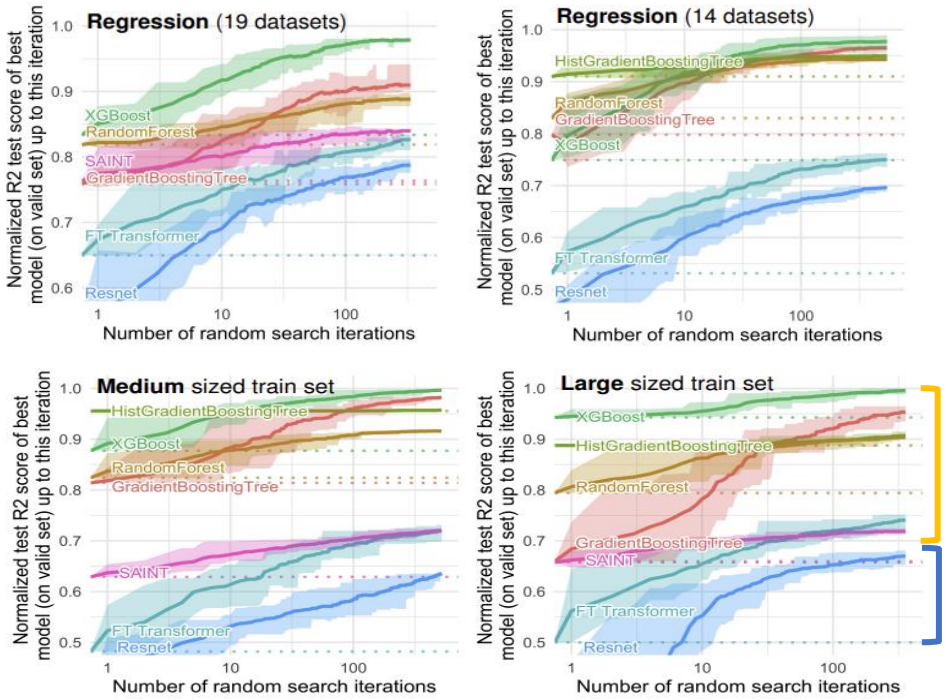


① 현재 시점에서 알 수 없는 미래의 관측 변수와 미래 시점의 값을 알 수 있는 변수 활용
→ Time Series Data에 대한 Prediction 성능을 높임

② 원하는 시점 단위를 학습하기 위해 부족한 컴퓨팅 파워,
시점을 특정 단위로 끊어서 예측할 경우에 성능 안좋음

학습하기 위한 컴퓨팅 파워의 부족하여 학습 시간이 오래 걸리며 Train MAE에 비해서 Vaild MAE가 낮았다

Tree 모델과 딥러닝



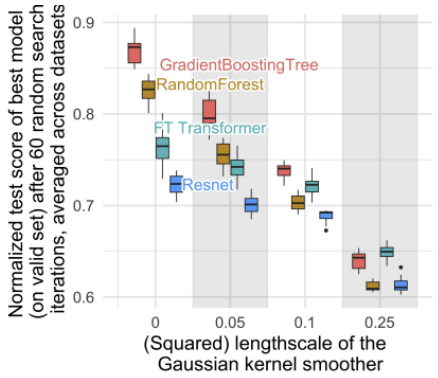
** 데이터 크기에 따른 R^2 score 측정 시,*
Tree 기반 모델과 딥러닝 모델 간 차이 증가

정형 데이터에 있어 딥러닝 모델보다 Tree – Based Model이 더 뛰어난 성능을 보임
또한, Numeric features만 사용하거나, Numeric features와 Categorical features 사용한 두 경우 모두
최적의 파라미터를 찾은 이후, Tree기반 모델이 딥러닝 모델에 비해서 훨씬 좋은 성능을 보임

Tree 모델과 딥러닝

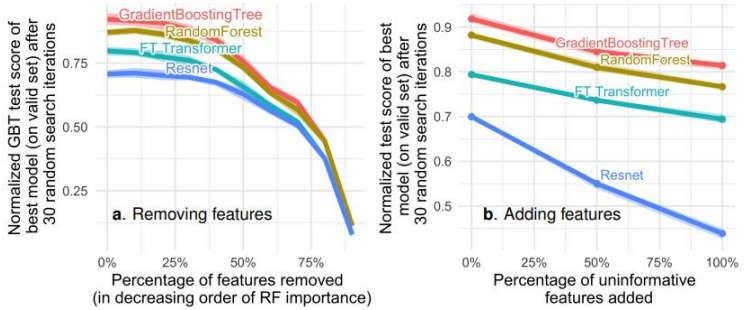
Tree 모델이 딥러닝 보다 뛰어났던 이유

불규칙한 패턴처리



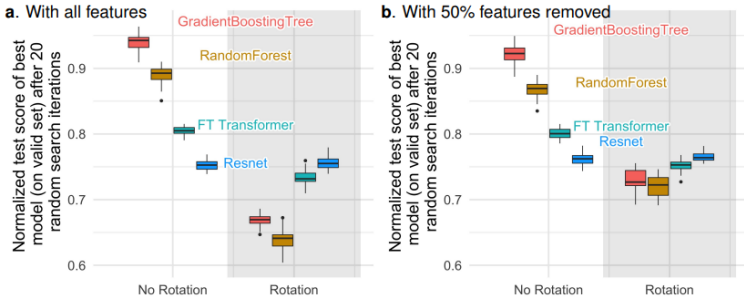
딥러닝 모델의 긴 smoothing 기간은
정형적 데이터를 너무 이른 시간에 smoothing하여
불규칙적인 패턴을 파악 어려움

Uninformative features의 존재와 모델의 강건함



트리 기반 모델들은 Uninformative features에
강건하지만, MLP 기반 모델은 그러지 못하여
이 변수를 제거하면 성능이 감소

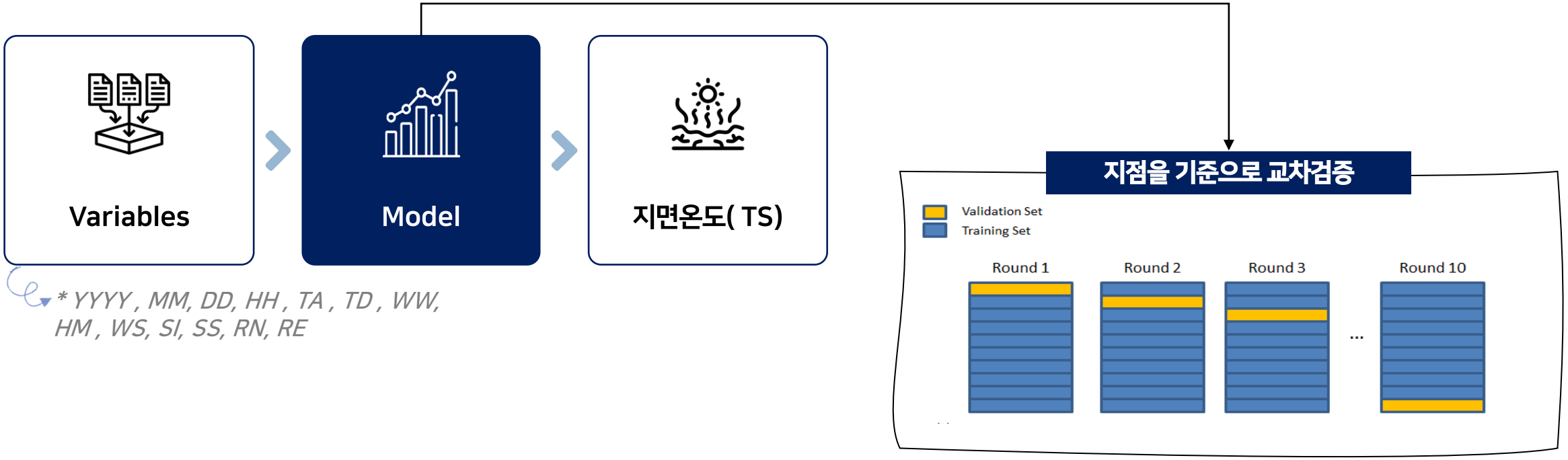
정형데이터의 non rotationally invariant



회전 처리의 진행 여부에 따른 성능을 비교해본 결과
ResNet은 정형 데이터에 부적합

Tree 모델은 딥러닝 모형에 ① 불규칙 패턴 파악 ② 유의하지 않은 변수를 포함한 예측 ③ 데이터 변형 측면에 비교우위를 가짐

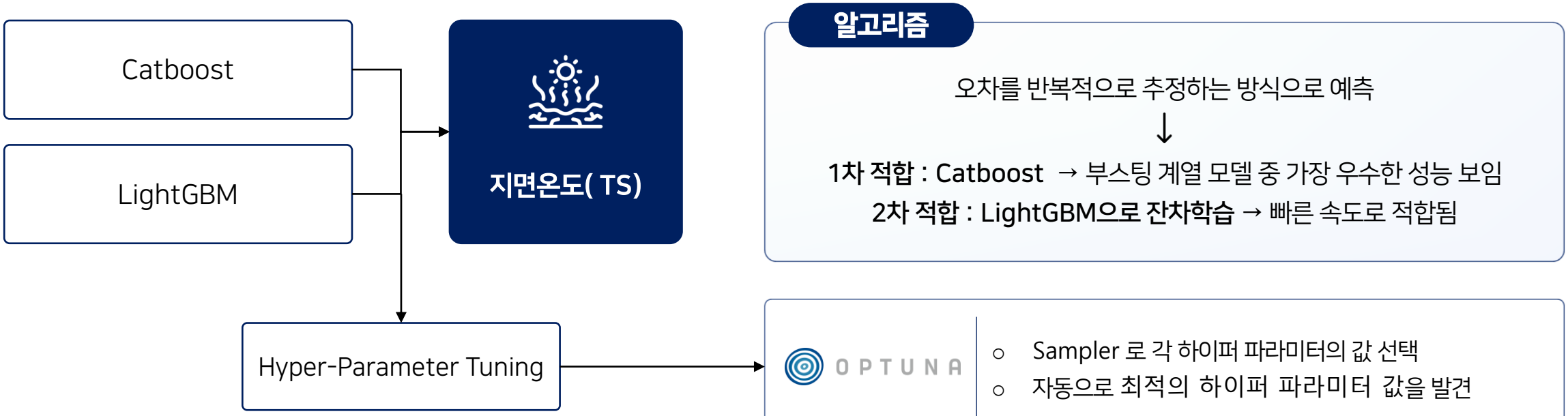
Catboost Regressor + LightGBM



분석과제: 새로운 관측지점에 관한 계절별 지면 온도 예측

지점에 관한 정보는 모델링에 직접적으로 반영하기 어렵기 때문에 **지점별로 데이터를 분리한 10개의 폴드**를 이용해 교차검증 실시

Catboost Regressor + LightGBM

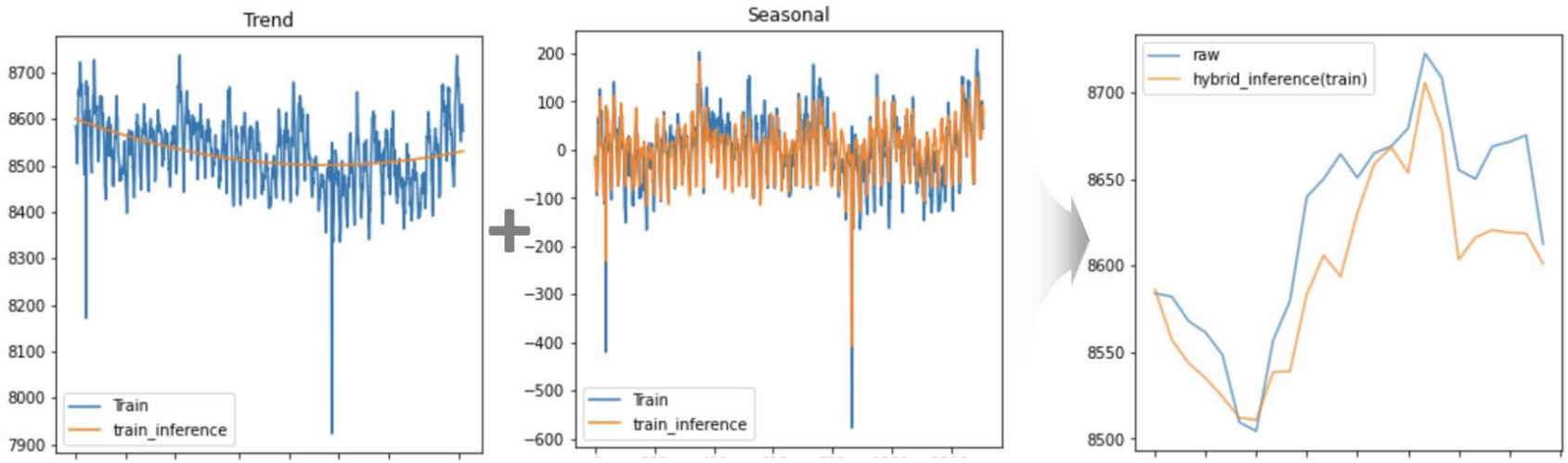


Python의 optuna 모듈을 사용해

지점별 CV를 통해 구한 **MAE의 평균값을 최소화하는 방향**으로 모델의 최적화를 진행

Stacked Hybrid Model

Time Series Component = Trend + Seasonality + Cycle + Error



시계열 성분은 **추세, 계절성, 주기**로 구성되어 있으며,

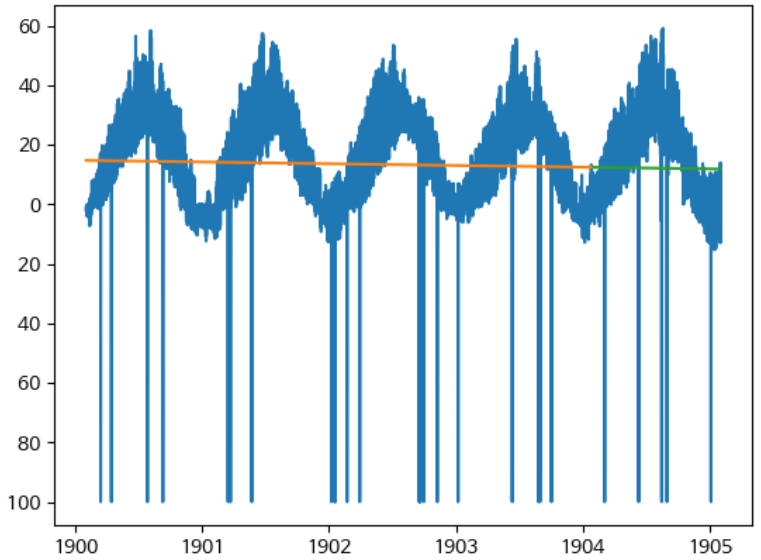
이 3가지 요소를 제거 후, 잔차만이 존재하는 백색잡음을 만들어 분석을 진행해야 함

↓

Stacked Hybrid model은 시계열 성분 각각을 모델링한 후, **성분 예측 값의 합**을 통해 시계열 성분 도출

Stacked Hybrid Model

10개 지점별 시작화 결과

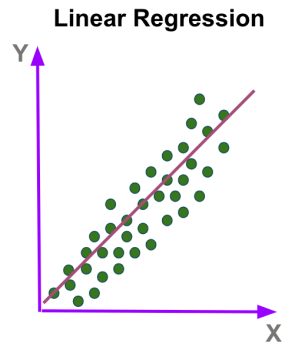


Stn1 지면온도 시각화

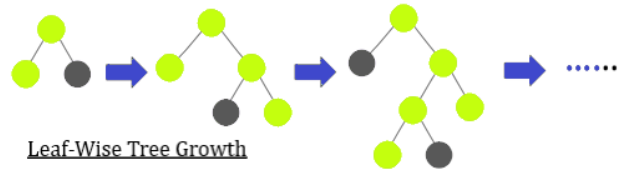
미세한 1차 추세 및 강한 계절성 관측

주기는 존재하지 않는 것으로 판단

추정 과정



1차 추세

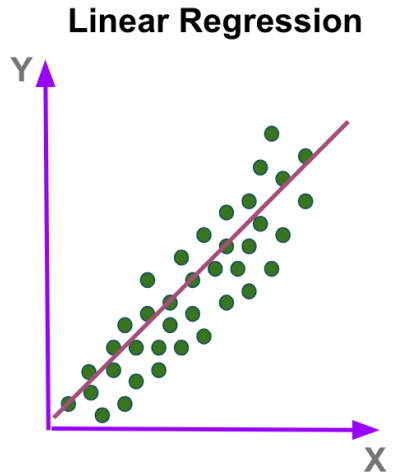


계절성

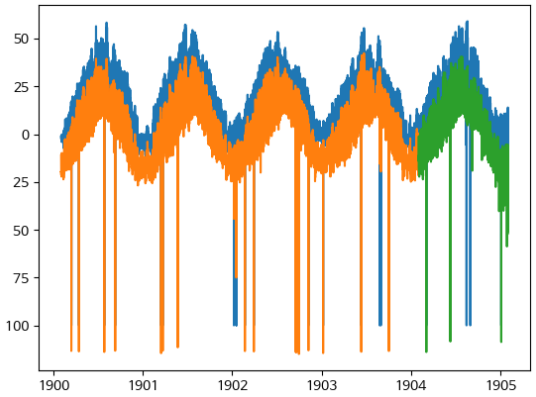
1차 추세는 단변량 선형회귀,

계절성은 XGB 및 LGBM 모델을 통해 추정

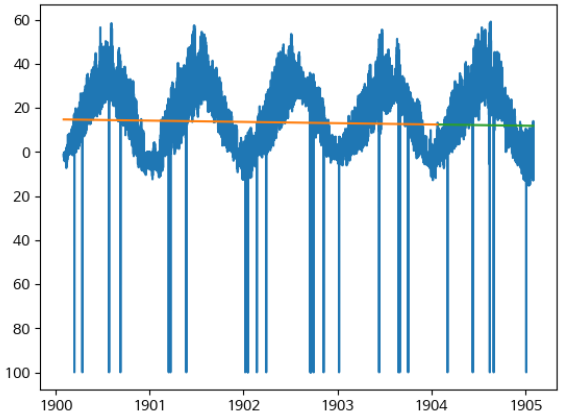
Stacked Hybrid Model



1차 추세 예측값



추세를 제거 후 LGBM을 통한
계절성 예측값



시계열 성분

① 지면온도 시계열 성분에서 회귀모형을 통한 1차 추세 예측 값 제거
② 계절성 존재 데이터에 대한 XGB과 LGBM 학습
③ **추세 예측 값과 계절성 예측 값의 합**을 통해 최종 지면온도 예측 값 산출

Stacked Hybrid Model

가을 - XGBoost

모델 자체에 과적합 규제 기능으로
강한 내구성을 지니며
내부적으로 결측치를 처리해줌

겨울 - LGBM

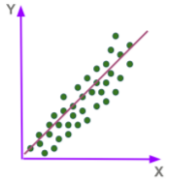
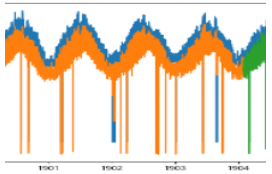
겨울 데이터에 **이상치 빈번히 발생**
많은 튜닝 과정 요구
↓
빠른 학습 속도 필요

지면온도(TS)

트리 모델은 이상치와 결측치에 강건하나,
계절 특성상 이상치의 빈번한 발생 및 기타 계절별 요인 때문에 겨울과 가을에 각각 다른 모델 사용

Stacked Hybrid Model

10개 지점별 Stacked Hybrid Model

( 1차 추세 예측 값 +  XGBoost / LGBM

OPTUNA

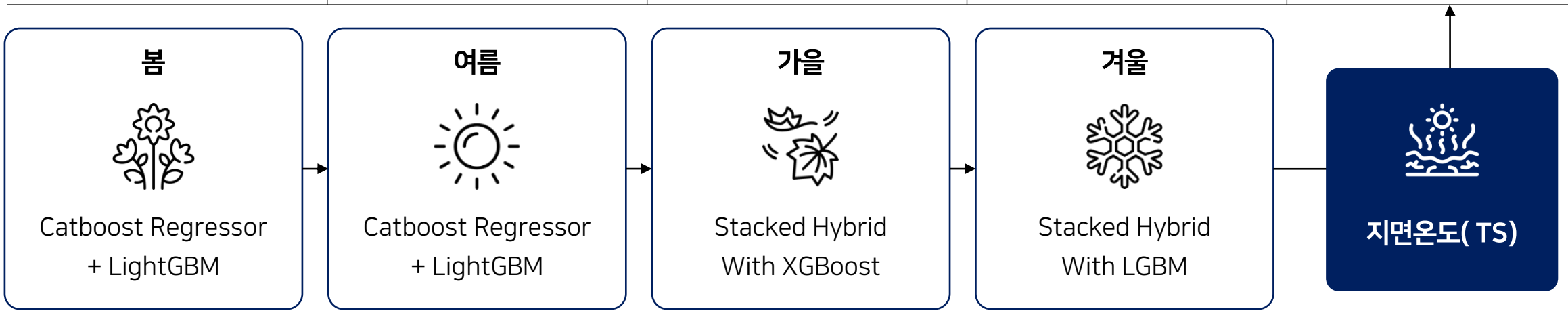
× 10) ÷ 10 = 지면온도(TS)

XGB/LGBM의 경우, 예측 성능 향상을 위해 베이지안 최적화 프레임워크 OPTUNA를 통한 하이퍼 파라미터 튜닝 진행

Soft Voting을 통해 10개 지점별 예측 값 앙상블 → 최종 3개 지역 예측 값 산출

최종 모델 선정

봄 MAE	여름 MAE	가을 MAE	겨울 MAE	AVG MAE
1.801	2.105	1.719	1.852	1.869



최종적으로 사용한 모델은 봄/ 여름에 Catboost Regressor + LightGBM이고,
가을에는 Stacked Hybrid with XGBoost, 겨울에는 Stacked Hybrid with LGBM이다.
최종 MAE: 1.869

활용방안 및 기대효과

농업에서의 활용

① 적절한 지면온도를 유지해야 하는 농작물의 특성을 고려하여
적절한 농작물의 수확방식과 시기를 정할 수 있음

② 여름철 야외에서 일하는 농민에게
폭염 알림을 전송하여 인명 피해를 줄일 수 있음

건설업에서의 활용

① 하절기 높은 지면온도로 인한 특정 화학물질이 포함된
건설 자재의 화학적 변화를 예측하며 위험을 사전에 방지할 수 있음

② 지면온도 예측을 통해 토양의 균열 발생 지역 등을
미리 예측하며 **안전하게 건설 지역을 선정 및 보완**할 수 있음

③ 여름철 야외 근로자들에게
폭염 알림을 전송하면서 인명 피해를 줄일 수 있음

활용방안 및 기대효과

에너지 생산 소비

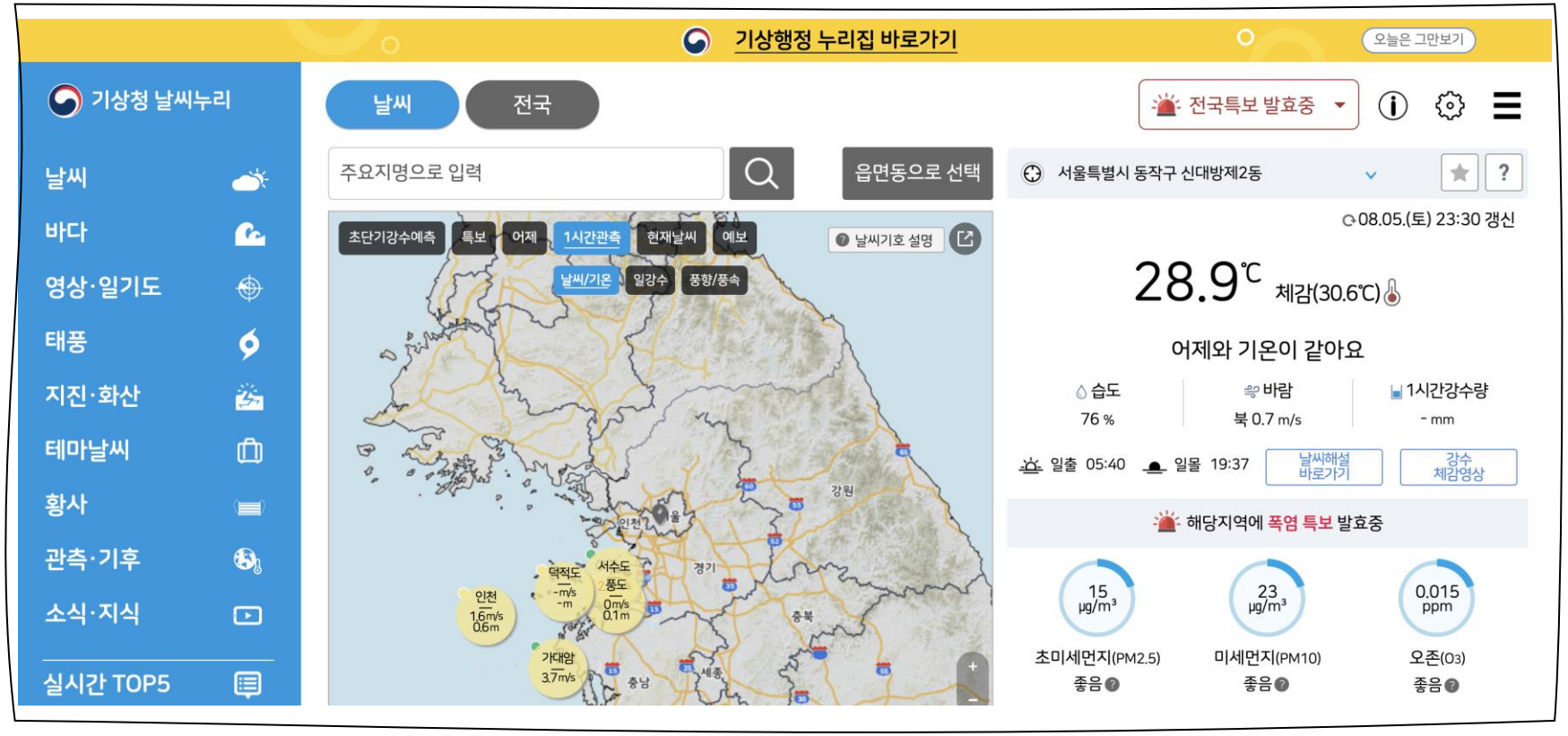
① 여름철 지면온도 상승으로 인해 **에너지 소비량**이
예상 정도 보다 높게 발생할 수 있음을 미리 예측하여 대비할 수 있음

② 지역별 지면 온도 예측에 따라
지열과 같은 **재생 에너지 최적화에 기여**할 수 있음

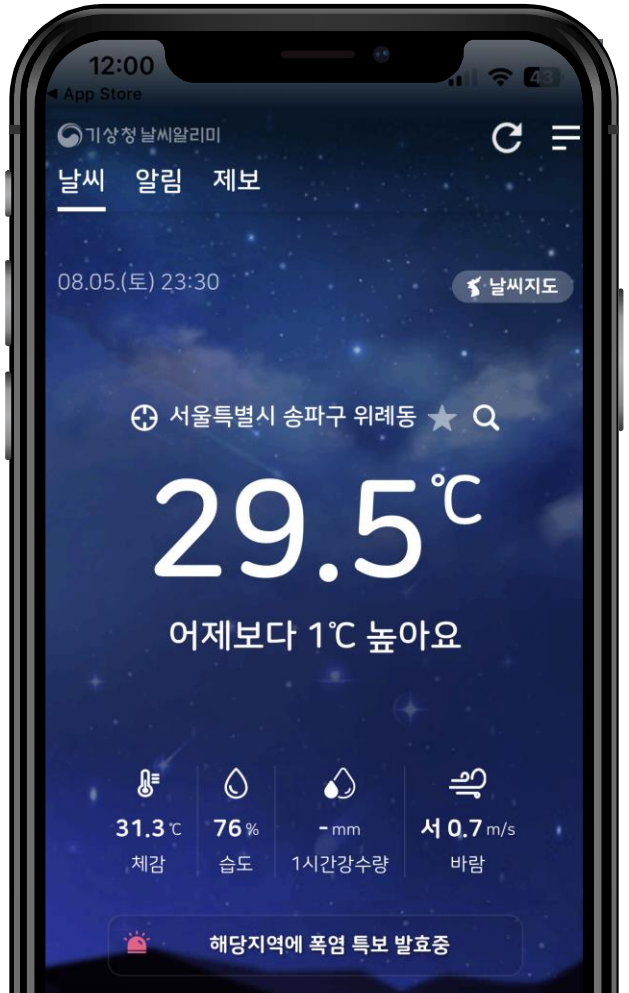
겨울철 결빙 방지

겨울철 **결빙 지역을 정확하게 예측**하고
사전에 이로 인한 자동차 사고, 인명 사고 등을 대비하면서
안전사고를 획기적으로 줄일 수 있음

활용방안 및 기대효과



** 기상청 웹페이지 & 앱*



실행방안

지면온도를 제공하고 있지 않아 사람들은 실제 온도,
체감 온도, 지면온도의 차이를 비교할 수 없기 때문에
기상청 앱과 웹페이지에 지면온도에 대한 내용 추가

기대효과

지면 온도 예측 정보를 제공함으로써
① 더위, 추위 등을 대비하는데 도움을 얻을 수 있음
② 인적· 물적 피해를 최소화 할 수 있음

참고문헌

[1] 기상청, 여름철 폭염, 이래서 위험합니다!
[2] 기상청, 지상:종관기상관측(ASOS) 자료 https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36
[3] 김민경, [날씨] 예측 힘든 야행성 게릴라 장마... 당분간 내륙은 폭염, YTN. (2023.06.30.), https://www.ytn.co.kr/_ln/0108_202306301254489130
[4] 박연선, KBS뉴스, (2022.12.07) 출근길 살얼음 사고 6명 사상…"지면 온도 더 낮아 유의" https://news.kbs.co.kr/news/view.do?ncd=5618929
[5] Scikit-learn, Imputing missing values with variants of IterativeImputer
https://scikit-learn.org/stable/auto_examples/impute/plot_iterative_imputer_variants_comparison.html
[6] Léo Grinsztajn, Edouard Oyallon, Gaël Varoquaux , Why do tree-based models still outperform deep learning on typical tabular data? (2022)

분석도구

jupyter python R Studio