기상청

기상에 따른 계절별 지면온도 산출 기술 개발
: 최종 MAE 1.869

1. 분석 배경 및 목표

지면 온도는 지표면 부근에서 측정한 온도를 나타내며, 여름철 일사, 겨울철 결빙 등 국민 실생활과 농업 등 다양한 응용 분야에서 활용되기 때문에 중요한 지표이다. 지면 온도는 시간, 기온, 계절 등 다양한 기상변수에 영향을 받는데, 특히 우리나라의 경우 사계절이 뚜렷하게 구분된다. 최근 여름철 야행성 게릴라 장마 및 폭염이 기승을 부리고 있는 상황에서, 시공간적으로 이를 예측하기 위한 지면 온도의 수요는 점차 증가하고 있다. 그러나 지면 온도 관측을 위해 진행되는 종관기상관측의 경우, 국내 103개 지점에서의 관측값만 알 수 있으며, 다른 지역의 관측값을 얻기 위한 물적, 인적 투입이 불가피한 상황이다.

따라서, 기존의 기상 빅데이터를 활용하는 것과 더불어 계절별 기후 특성을 반영하여 유의미한 예측값을 제공하는 계절별 지면 온도 산출 기술의 개발이 필요하다. 지면 온도 산출 기술의 개발을 위해, 주어진 10개 지점에 대한 기상 관측 데이터를 활용해 새로운 지점의 지면 온도를 예측하는 모델링을 진행하였다.

2. 데이터 정의

지면 온도 예측 모델링을 위한 데이터는 기상청에서 제공받은 것으로, 10개의 기상 관측 지점에 대한 다년간의 기상 관측 데이터이다.

테이블명	내용	테이블명	내용
YYYY	년도	HM	1시간 평균 상대습도
MMDDHH	월/일/시간	WS	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

<표1> 지점별 기상 관측데이터 항목별 구성

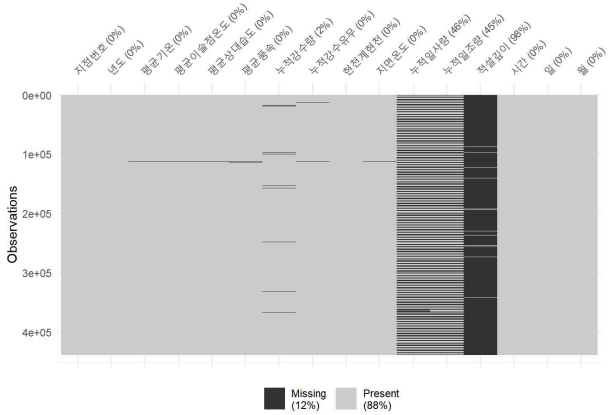
이때, 주어진 데이터에는 기상청 제공 자료에서 제시된 바와 같이 -99,-99.9,-999와 같은 결측치가 다수 존재했고, 추후 결측치 처리를 위해 NA값으로 변경했다. 특히, 정확한 모델링을 위해 종속변수인 지면 온도(TS)에 결측치가 존재하는 경우 해당 행을 제거하여 학습에서 제외했다.

새로운 관측지점에 관한 계절별 지면 온도 예측이라는 최종 분석과제는 고려하여, 첫째, MMDDHH를 기준으로 계절별로 데이터 분리하였다. 둘째, 데이터의 관측 지점에 대한 정보와 계절적 연속성을 고려한 다양한 교차 검증 방식을 채택했다. 또한, 학습데이터와 평가에 사용되는 검증 데이터(Test data)의 행 개수 비율을 고려해 8:2 비율로 검증을 진행했다.

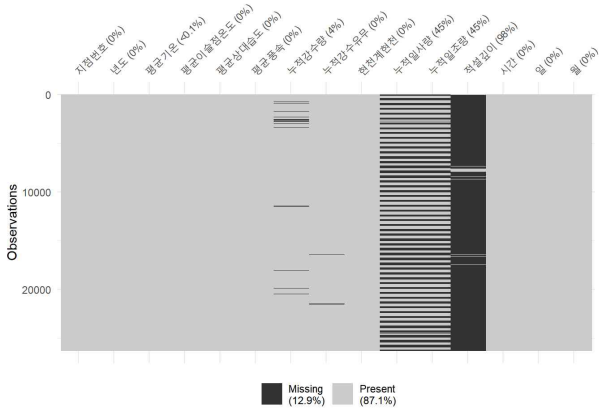
3. 탐색적 자료 분석(EDA)를 통한 전처리(Pre-Processing)

통계량 및 시각화 등의 방법을 통해 EDA를 진행했으며, 이를 기반으로 데이터의 전처리를 진행했다.

3.1 결측값(Missing Value) 처리

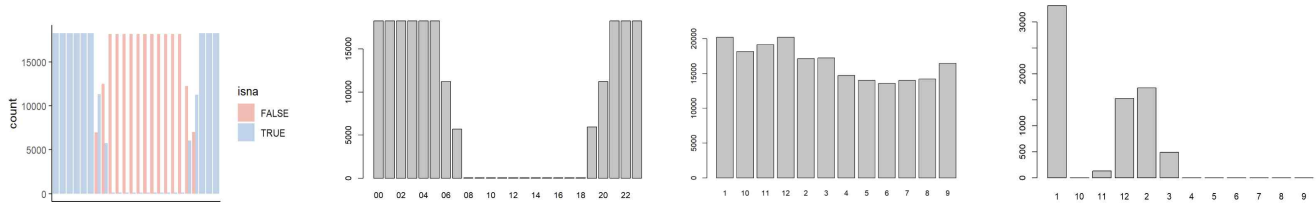


<그림1> 학습데이터의 결측치 시각화



<그림2> 검증데이터의 결측치 시각화

R의 naniar 패키지와 Python의 msno 모듈을 통해 결측치를 시각화해 본 결과, 종속변수를 포함한 수치형 변수에서 결측이 발생했고, 그중 누적 일사량(SI), 누적 일조량(SS), 적설 깊이(SN)의 경우 그 결측치의 비율이 40%가 넘는 상당한 비중을 차지했다. 이러한 결측치에 발생 패턴이 존재하는지를 확인하기 위해 ① 결측치인 경우 ② 결측치가 아닌 경우로 나누어서 시각화를 진행했다.



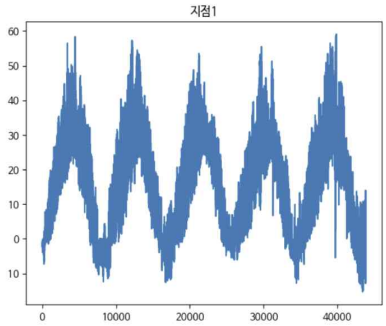
<그림3> 시간/월에 따른 1시간 누적 일조량 결측치 시각화

<그림4> 월별 적설깊이 데이터 시각화

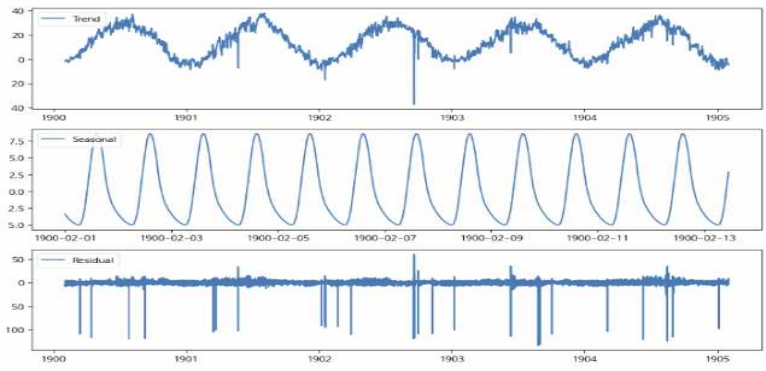
결측값이 나타난 변수는 총 9개로, 결측치 비율이 낮은 변수의 경우, 결측치 발생 패턴이 뚜렷하게 나타나지 않았다. 반면, 결측치의 비율이 높은 변수 중 하나인 1시간 누적 일사량(SI)와 일조량(SS)의 경우, 밤 시간대에 주로 결측값이 발생한 것을 확인했다. 결측치 발생 비율이 가장 높았던 적설깊이(SN)은 눈이 오지 않는 계절의 경우 모두 결측치 처리가 되어 결측치 비중이 높았던 것으로 판단된다. 이를 통해, 결측값의 발생 패턴이 존재함(MNAR, Missing Not At Random)을 확인했다. 결측치 비율이 높았던 SI, SS, SN의 결측치는 학습데이터와 검증데이터 모두 0으로 대체하였다.

나머지 변수의 경우, 상관관계가 높은 변수 조합을 고려하여 Python의 Scikit-learn의 Iterative Imputer의 Bayesian Ridge를 이용해 결측치 보간을 진행했다. 검증데이터에도 결측이 발생했는데, 이때, Data Leakage를 방지하기 위해 Iterative Imputer를 적합시킬 때는 학습데이터의 정보만 이용하였다. 하지만, 종속변수인 TS의 결측값을 보간할 경우 데이터의 왜곡이 발생할 것이라 판단하여, TS에서 결측값이 발생한 행을 삭제했다.

3.2 시계열 특성 시각화 (TS Plot)



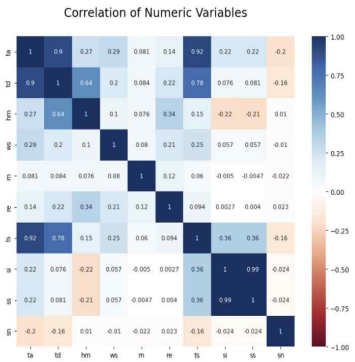
<그림5> 지점의 지면온도 변수의 시계열 플랏



<그림6> 시계열적 요소를 분해하여 시각화한 결과

모든 10개 지점의 지면온도에 대해 시계열 플랏을 그려 시각화해본 결과, 추세 및 계절성이 존재하는 비정상 시계열임을 확인했다. <그림6>은 시계열적 요소를 분해(decompose)한 결과로, 경미한 1차 추세(trend), 강한 계절성(Seasonality)이 존재하며, 분해 이후 잔차의 분포가 랜덤한 것으로 보아 순환요인(cycle)은 존재하지 않는 것을 확인했다.

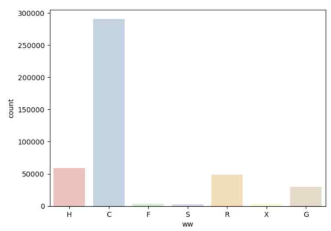
3.3 수치형 변수 간 상관관계 파악

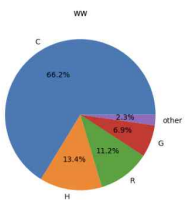


<그림7> 수치형 변수 간 상관관계 시각화

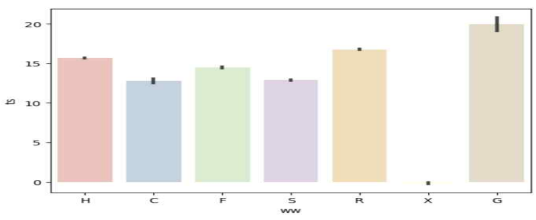
변수 간 선형관계를 보기 위해, 피어슨 상관관계를 구해 히트맵으로 시각화했다. 평균 기온(TA), 평균 이슬점온도(TD), 평균 지면온도(TS) 간에서 높은 상관관계를 가지는데, 이는 지면온도(TS)가 높을수록, 평균 기온(TA)과 평균 이슬점온도(TD)도 높아짐을 의미한다. 또한, 누적 일사량(SI), 누적 일조량(SS) 변수 간에도 높은 상관관계를 보인다. 즉, 두 변수의 경향성이 매우 비슷하다고 볼 수 있다.

3.4 범주형 / 수치형 변수 시각화





<그림8> 현천계 현천(WW) 의 유형별 개수와 비중

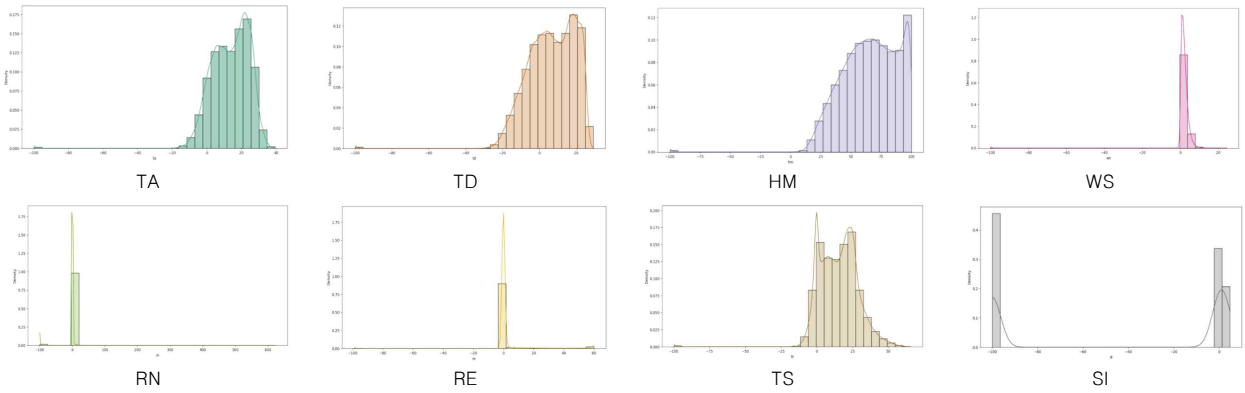


<그림9> 현천계 현천(WW) 유형별 평균 지면온도(TS)

<그림8>을 통해 범주형 변수인 현천계 현천(WW)의 유형별 개수를 확인해 보았을 때, 7개 유형 중 맑음(C)이 가장 높은 비율을 차지했다. 그 다음, 박무(H), 비(R), 연무(G) 순으로 비중이 이어졌다. 안개(F), 눈(S), 모름(X)는 비슷한 비율을 가진다.

<그림9>는 현천계 현천(WW) 유형별 평균 지면온도(TS)를 나타낸다. 연무(G)일 때 평균 지면

온도(TS)가 가장 높은 것으로 확인되었다. 그 다음으로 비(R), 박무(H), 안개(F)순으로 평균 지면온도(TS)가 높게 나타났으며, 맑음(C)과 눈(S)의 평균 지면온도(TS)는 비슷한 값을 가진다. 신뢰구간(검정선)을 봤을 때, 연무(G)가 가장 긴 것으로, 신뢰도가 떨어진다는 것을 확인할 수 있다.



<그림10> 수치형 변수 분포 시각화

<그림10>을 통해 평균 기온(TA), 평균 이슬점온도(TD), 평균 상대습도(HM)의 분포 형태가 비슷함을 확인했다. 반면, 평균 풍속(WS), 누적 강수량(RN), 누적 강수유무(RE)는 치우친 분포를 보인다.

3.5 이상값(Outlier) 처리

앞선, 수치형 변수 분포 시각화를 통해 이상치의 존재를 확인했다. 이상치 처리에 앞서, 주어진 예측 모델링의 평가지표인 MAE(평균 절대 오차)는 이상치에 강건하다는 특징이 있으며, 지면온도 추정이라는 데이터 분석 목적을 고려한다면 데이터의 이상치 역시 의미 있을 것이라 판단했다. 대신, 데이터의 중앙값이 0, IQR(Inter Quantile Range)=1이 되도록 스케일링하는 기법인 로버스트 스케일링(Robust Scaling)을 진행하여, 이상치의 영향을 최소화했다.

3.6 현천계 현천(WW) Encoding

7개의 범주(S: 눈, R: 비, F: 안개, H: 박무, G: 연무, C: 맑음, X: 모름)를 가지고 있는 현천계 현천(WW)의 경우 모델링에 따라 Label Encoding 또는 One-Hot Encoding을 진행하여 수치형 변수로 변경해 주었다.

4. 모델링(Modeling)

4.1 Catboost Regressor 기반 커스텀 모델

<그림11> 지점을 기준으로 교차검증 진행

새로운 관측지점에 대한 지면온도 예측이라는 분석과제를 고려할 때, 지점에 관한 정보를 모델링에 직접적으로 반영하는 것에는 어려움이 있었다. 따라서, 지점별로 데이터를 분리한 10개의 폴드를 이용해 모델에 대한 교차 검증을 실시했다.

교차검증을 통해 여러 모델(선형 모델 – Lasso, Ridge, Bayesian Ridge, Elastic-Net, 부스팅 모델 – LGBM, XGBoost, LGBM,Catboost 등)의 성능을 비교해 보았을 때, 오차를 반복적으로 추정하는 방식을 통해 학습하는 HistGradientBoosting Regressor의 MAE가 비교적 낮았다. 이에 착안하여 잔차를 반복적으로 추정하는 커스텀 모델을 제작했다. 1차 적합에는 부스팅 계열 모델 중 우수한 성능을

보이는 Catboost Regressor를 사용했고, 잔차를 학습하는 2차 적합에는 빠른 속도로 적합되는 LGBM을 이용했다.

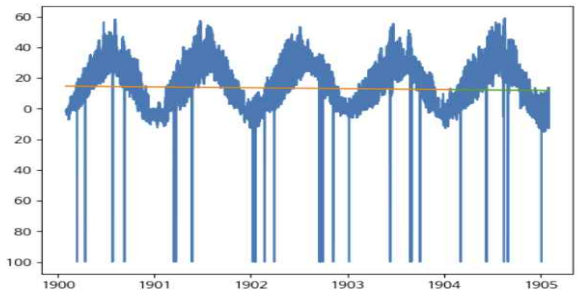
모델의 성능 향상을 위한 하이퍼 파라미터 튜닝에는 Python의 optuna 모듈을 사용했다. 해당 모듈을 활용해 지점별 CV를 통해 구한 MAE의 평균값을 최소화하는 방향으로 모델의 최적화를 진행했다. optuna를 통해 얻은 Catboost, LGBM의 최적의 하이퍼파라미터는 다음과 같다.

봄	Catboost	learning_rate	0.04158500359538149
		depth	12
		iterations	419
		has_time	True
	LGBM	max_depth	15
		learning_rate	0.09950693202165563
		n_estimators	902
		min_child_samples	59
		subsample	0.8052368021730489
여름	Catboost	learning_rate	0.12515542897654602
		depth	13
		iterations	401
		has_time	True
	LGBM	max_depth	14
		learning_rate	0.09990601824979606
		n_estimators	924
		min_child_samples	74
		subsample	0.30422223766873524

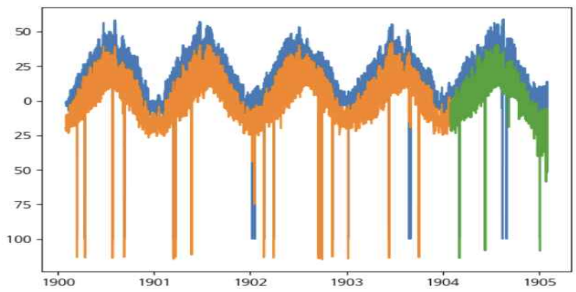
<표2> optuna를 통해 얻은 커스텀 모델의 파라미터

4.2 Stacked Hybrid

$$\textit{Time Series Component} = \textit{Trend} + \textit{Seasons} + \textit{Cycles} + \textit{Error}$$



<그림 12> 1차 추세 예측



<그림 13> LGBM을 통한 계절성 데이터 학습 및 예측

비정상 시계열은 추세, 계절성, 순환과 잔차로 구성되어있다. 적절한 시계열 예측을 위해서는 잔차를 제외한 나머지 요소들을 제거하여, 정상 시계열인 백색잡음으로 만들어야 하므로 차분 등의 과정이 필수적이다. 따라서, 본 분석에서는 해당 과정을 모델링을 통해 해결한 Stacked Hybrid 모형을 채택하였다.

Stacked Hybrid 모형은 우선 추세를 선형 회귀 모델로 학습시켜, 해당 추세 추정값을 훈련 데이터에서 제거하고 이후 계절성만 존재하는 훈련 데이터를 트리 모형 등으로 학습시켜 예측값을 산출한다. 이후, 추세 추정값과 계절성만 존재하는 데이터로 학습한 예측값을 합하여 최종 예측값을 구한다.

<그림12>에서 볼 수 있듯이, EDA를 통한 시계열 분해 결과, 미세한 1차 추세가 관측되었으므로, 1차 선형 회귀 모형을 사용했다. 계절성만 존재하는 데이터의 경우, 계절에 따라 트리 기반 모델인 XGBoost와 LGBM을 사용했다. 가을의 경우 XGBoost(Extreme Gradient Boosting)의 성능이 좋아 이를 사용했다. 겨울의 경우 LGBM을 사용했는데, 이는 겨울 데이터 특성상 이상치 발생이 빈번함에 기인한

다. 즉, 더 많은 튜닝 과정을 요구하여 빠른 모델의 학습 속도가 필요했다. 또한, 학습 방향을 지정해주는 목적함수로써 이상치에 강건한 Huber loss를 적용시켜 통해 이상치의 효과를 최소화하고자 했다.

이때, 사용한 트리 모델의 하이퍼 파라미터 튜닝을 위해, 베이지안 최적화 프레임워크인 optuna를 활용하여 모델을 튜닝했다. 또한, 모델 튜닝 과정에서 과적합 방지를 위해 학습데이터를 8:2 비율로 나누어 Train:Validation Data Set을 각각 구성하여 모형을 학습 및 적합했다.

5. 모델을 이용한 예측 (Prediction)

봄·여름의 경우 Catboost Regressor 기반 커스텀 모델을, 가을· 겨울의 경우 Stacked Hybrid 모델을 이용해 최종적으로 지면온도 예측을 진행했다. 각 계절의 모델을 통해 얻은 검증 데이터의 최종 예측 성능은 <표3>과 같다.

MAE				
SPRING	SUMMER	AUTUMN	WINTER	AVG
1.801	2.105	1.719	1.852	1.869

<표3> 최종 예측 성능

6. 활용 방안 및 기대효과

지면온도를 정확하게 예측하는 것은 실생활과 응용 분야에서 매우 중요하다. 그러나 기상자료 수집 과정에서 결측값과 이상치가 많이 존재하며, 우리나라의 지리적 위치상 사계절의 기상 관측값은 명확하게 차이가 있다. 프로젝트를 통해 이러한 기상 자료 특성을 모두 고려하여 데이터를 계절별로 나누어 각 계절에 부합하는 모델을 개발했다. 그 결과, 단일 모델보다 정확한 지면온도 추정값을 산출할 수 있었다. 이를 통해 지면온도 예측을 위한 불필요한 인적 · 물적 낭비를 줄이고, 기상 관측이 불가능한 이상 기후를 사전적으로 예측하여 이로 인한 기후 재난적 상황을 대비하고 예방할 수 있을 것이다.

References

1. 김민경, [날씨] 예측 힘든 야행성 게릴라 장마... 당분간 내륙은 폭염, YTN. (2023.06.30.), https://www.ytn.co.kr/_ln/0108_202306301254489130
2. 기상청, 지상:종관기상관측(ASOS) 자료 https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36