기상에 따른 계절별 지면온도 산출 모델 개발

분야	참가번호	팀명
생활안전	230064	통상실의 전사들

I. 분석 배경 및 목표

지면온도는 폭염이나 한파에 효과적으로 대응할 수 있는 필수 지표이며, 식량 안보, 농업 생산성, 수자원 관리, 관광 산업 등 다양한 분야에서 중추적인 역할을 하는 기상 요소다. 국민 생활에 밀접하게 연결되어 있기 때문에 지면온도 정보에 대한 수요는 계속해서 증가하는 추세이다. 그러나 현재 지면온도를 측정하는 위치가 상대적으로 부족하여, 많은 관측 공백 지역이 존재한다. 이로 인해 현장에서 필요로 하는 정확한 지면온도 정보 제공에 어려움을 겪고 있다. 위 문제를 해결하기 위한 방안으로, 기상자료를 활용하여 지면온도를 보다 정확하고 세밀하게 예측하는 모델을 개발했다. 새로운 모델을 통해 기후 변화에 대응하는 능력을 강화하고, 국민의 안전에 기여하고자 한다.

II. 데이터 정의 및 전처리

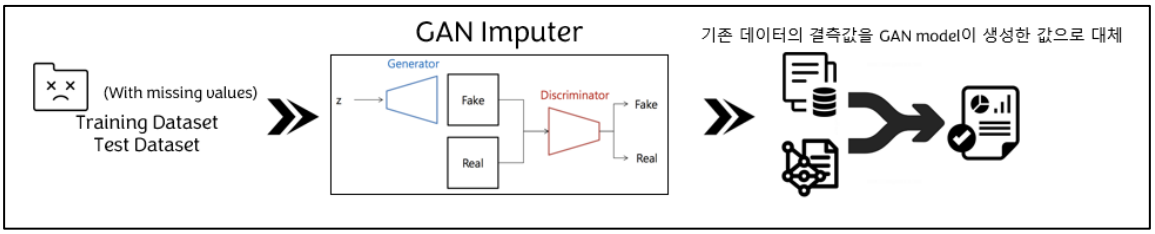
1. 데이터 정의 및 결측값 대체

데이터는 10개 area에 대하여 수집한 5년(A년~E년)의 기상 자료이다. 종속변수인 지표면 온도와 독립변수인 다양한 10가지의 기상지표로 구성되어 있는데, 서로 다른 지역에 대한 데이터이기 때문에 각 area별 지표가 동일한 분포를 가지고 있다고 보기 어렵다. <표1>를 통해 이를 직접적으로 확인할 수 있는데, area별로 기온 관련 지표들이 현저한 차이를 보이는 것을 파악할 수 있었다. 특히 area1과 10을 직접적으로 비교해 보면 동일한 특성을 가진 지역이라고 말할 수 없을 정도로 지표간 큰 차이를 보여주고 있기에, 이에 유의하여 모델링을 진행했다.

	area 1	area 2	area 3	area 4	area 5	area 6	area 7	area 8	area 9	area 10
봄-평균 기온	6.23	7.82	7.23	6.56	6.79	8.05	9.11	8.99	10.17	9.61
봄-기온 표준편차	7.79	6.29	7.04	6.14	6.95	7.16	6.79	6.51	5.26	5.29
봄-평균 일교차	21.51	17.08	17.87	14.19	18.48	19.74	19.15	18.34	14.75	14.15
봄-평균 최고 기온	17.56	17.22	16.95	14.38	16.83	18.67	19.47	19.23	17.51	16.84
봄-평균 최저 기온	-3.95	0.15	-0.91	0.19	-1.66	-1.07	0.32	0.89	2.77	2.69

<표1> Area별 봄 기온 관련 지표

또한 주어진 훈련 데이터에 다량의 결측값(-99, -99.9등으로 표시)이 존재하는 것을 확인할 수 있었다. 훈련 데이터뿐만 아니라 검증 데이터에도 결측값이 존재하기 때문에 올바른 모델링과 평가를 위해서는 적절한 data imputation 작업이 선행되어야 한다. 현재 데이터가 각 area에 대한 Multivariate time series data로 주어져 있고, 데이터가 가진 시계열 자료의 특성과 변수들 간의 관계를 반영하고자 각각의 area에 대하여 GAN(Generative Adversarial Network)을 사용한 imputation 모델을 구현하여 사용했다.



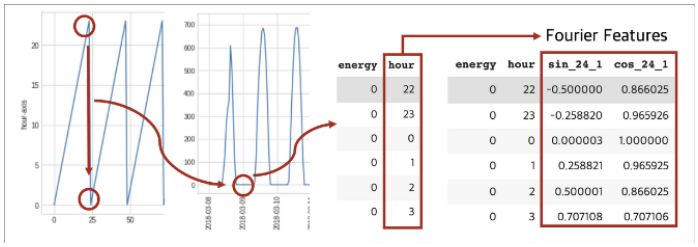
먼저 각 area별로 결측값의 위치를 찾아 마스킹 한 후, mean imputation을 수행하여 결측값을 채워주었다. 다음으로 GAN 아키텍처를 사용하여 LSTM 레이어와 dense 출력 레이어를 기반으로 한 생성자(Generator)와 CNN을 기반으로 한 판별자(Discriminator)를 구현하여 학습시킨 후, 초기에 마스킹 된 위치에 생성자가 만들어낸 imputed data를 집어넣는 방식으로 data imputation을 진행했다. 추가적으로 위의 imputation 과정을 하나의 파이프라인으로 구축하여, 데이터가 추가되거나 바뀌더라도 파이프라인을 활용하여 손쉬운 결측값 대체가 가능하도록 해 주었다.

2. Fourier / Wavelet Transformation

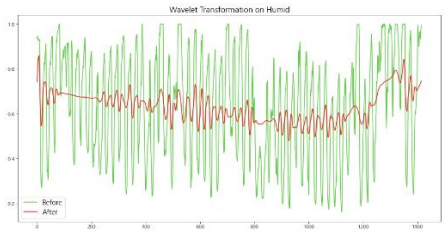
지면 온도는 하루 24시간을 기준으로 주기성을 지닌 시계열 데이터로 볼 수 있다. 머신러닝 모델이 주기성을 학습하기 위해 시간 순차성 특성을 지닌 시간(hour) 변수를 Fourier Bases를 이용하여 연속성을 가질 수 있도록 표현했다. 다음의 식을 이용하여 변환해 주었다. (그림1 참고)

$$\sin \text{hour} = \sin(2\pi h/24), \quad \cos \text{hour} = \cos(2\pi h/24) \ (h: \ hour)$$

또한 특정 독립변수의 노이즈를 제거하여 증감 추세와 같은 명확한 패턴만을 남기고자 Discrete Wavelet Transformation을 활용했는데, 특히 데이터 압축에 효과적이라는 db5 모 웨이블렛 함수를 사용했다. 결과적으로 노이즈가 있는 기존 변수를 사용할 때 보다 노이즈가 제거된 변수를 사용했을 때에 머신러닝 모델이 좀더 안정적으로 작동하는 것을 확인할 수 있었다. (그림2 참고)



<그림1>-Fourier transformation



<그림2>-Wavelet transformation

3. Month, Day 변수 변환

month와 day 변수를 묶어서 월별 특징을 파악하고자 했다. 이미 hour변수를 이용하여 주기성을 표현했고 더불어 day변수는 과도하게 세분화되어 지면온도 값을 오히려 제한시킬 수 있다고 판단했다. 따라서 day 변수를 week 변수로 변환해 주었다. 변환 후 month와 week 변수를 묶고 범주화 해 주었다.

4. Lag / Weighted moving average / 상호작용항(Interaction Term)

지면온도는 단순히 현재 기상지표에만 영향을 받는 것이 아니기 때문에 이전의 기상상황에 대한 고려가 수반되어야 한다. 따라서 필요한 각종 독립변수에 대해 Lag를 새로운 변수로 추가하거

나 이전 관측값들을 모아 만든 가중평균 등과 같은 변수를 생성하여 모델링에 사용했다.

또한 일사량, 강수, 기온 등과 같은 변수들은 독립적으로 변화하지 않고 서로 밀접한 연관관계를 가지기 때문에 각각의 계절에 맞는 특성을 적절하게 고려하여 새로운 상호작용항을 만들어 모델링 과정에서 변수로 활용했다.

5. Snow 변수 처리

눈이 오게되면 구름이 많이 끼게 되어 맑은 날에 비해 비교적 지표면이 가진 열을 덜 빼앗기게 된다. 즉, 지면이 복사냉각의 영향을 덜 받기 때문에 지면온도가 일정하게 유지되는 경향이 있는 것이다. 이러한 점을 고려하여 기온과 일사량 값을 적설량으로 나눠주고 주기성을 제거시켜 줌으로써 만약 눈이 오게 된다면 적설량으로 나눠준 기온과 일사량 값을 작아지게 하여 눈이 오는 경우에 일사량과 기온이 지면온도에 미치는 영향을 비교적 낮게 반영되도록 해 주었다.

6. 계절별 현천 범주화 방법

주어진 현천계 범주를 계절의 특성을 고려하여 <표2>와 같이 수정했다. 눈이 오는 봄과 겨울은 비와 눈을 합쳐서 하나의 범주로 만들고, 여름의 경우에는 강수의 영향을 많이 받는 부분을 고려했다. 가을은 강수가 많지 않고 일조량의 영향을 줄 수 있는 안개 범주만 독립적으로 분리해 주었다.

	봄	여름	가을	겨울
1	맑음	맑음	맑음+X	맑음
2	비+눈	비	안개	비+눈
3	나머지	나머지	나머지	나머지

<표2>-계절별 현천 범주화 방법

7. 기타 파생 변수

그 외에도 비가 내린 직후의 증발냉각 효과로 인해 지면온도가 감소하거나 눈으로 인한 담요효과로 지면온도가 유지되는 경향을 반영하기 위하여 비와 눈이 내린 후 24시간까지의 경과 시간을 나타내는 변수인 after_rain, after_snow를 추가해 주었다. 또한 기온과 이슬점 온도의 차이를 나타내는 temp_diff 변수나 누적 강수나 적설량등과 같은 추가적인 파생변수도 활용했다.

III. 모델링 과정

1. 개요

데이터를 각 계절별로 나누어 계절별 모델링을 진행했다. 회귀 모델은 LightGBM이나 CatBoost 같은 부스팅 계열의 모델을 사용했고, 각 계절의 특성에 맞는 Lag 개수, 상호작용항, 파생변수 등을 다르게 설정하여 모델의 성능을 높이고자 했다. 주어진 데이터에서 시계열 자료의 특성을 고려하여 A~D년의 데이터를 training data로, E년의 데이터를 test data로 고정하고, training data를 통해 만들어진 모형을 test data로 검증하여 가장 좋은 test MAE를 가지는 변수들과 모형을 최종적으로 선택했다. 학습의 안정성을 위하여 Cross validation 과정에서 초기 split에 최소한 1년치의

데이터가 모두 들어가도록 time series split을 적용해 주었고, Bayesian optimization을 사용하여 하이퍼파라미터 튜닝을 진행했다.

계절의 특성에 따라 각 area별로 기상 분포의 차이가 큰 경우에는 각각의 Area에 대하여 개별적인 회귀 모델링을 진행한 후에 앙상블 기법을 사용하여 하나의 최종 모형을 만들어 주는 방식을 채택했고, Area에 관계없이 전체 기상 데이터를 활용하여 하나의 단일 모델을 만드는 것이 효과적인 계절에는 단일 모델을 만들어 최종 모델으로 사용했다.

2. 봄 모델링

봄은 비와 눈의 영향을 많이 받는 계절이므로 기상온도, 이슬점, 습도가 서로 밀접한 연관관계를 가지는 것을 확인해 해당 변수에 대한 상호작용항을 추가하여 사용했으며 기상온도, 이슬점, 적설량, 강수량, 일사량 변수에 대해 시간에 따라 반비례하게 부여한 가중치를 사용한 가중 평균(3시간)을 구해 변수로 사용했다. 또한 기상온도, 이슬점, 습도 변수에 대해서는 lag 변수(1,2,12,24시간)를 새로 정의해 주었다. 이때, 시간대를 shifting 하여 lag 변수와 가중 평균 변수를 생성했기 때문에 그 시간대(초기 24개 데이터) 만큼 초기 데이터에 결측값이 생기게 된다. 따라서 lag나 가중 평균 변수를 사용하지 않은 앙상블 모형을 추가로 학습하여 초기 데이터의 예측에 사용했다.

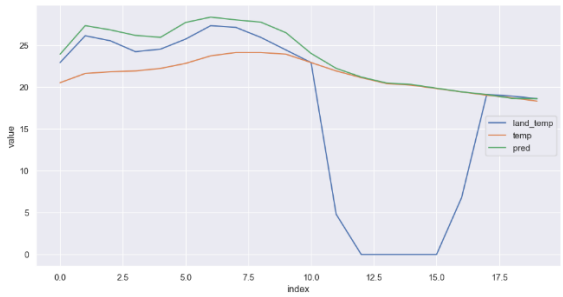
봄 모델링의 최종 모델로 선택된 CatBoost 모델의 경우 지면온도를 예측하는데 일사량과 기상온도의 영향을 많이 받기 때문에 습도와 이슬점 변수는 노이즈를 제거하여 증감 추이와 패턴 만을 나타낼 수 있도록 2-2)의 wavelet 변환을 이용했다. 초봄(2월) 에 눈이 오는 경우 보온효과가 일어나는 현상을 반영해주기 위해 눈이 온 시간대에 대해 앞서 설명한 2-5) 전처리 과정을 적용했다. 또한 2월의 경우 윤년에 의해 년도별로 데이터 다르기 때문에 area별로 time series split을 사용할 경우 A년도 4월의 데이터가 train set에 포함되지 않고 validation set에 포함되는 경우가 생기게 된다. 따라서 group time series split 방법을 사용하여 1개 년의 데이터가 학습하는데 전부 사용될 수 있도록 했다. 마지막으로 Bayesian Optimization (Gaussian Process) 방법을 이용해 area별 catboost 모델의 최적 hyperparameter를 찾아 주었다.

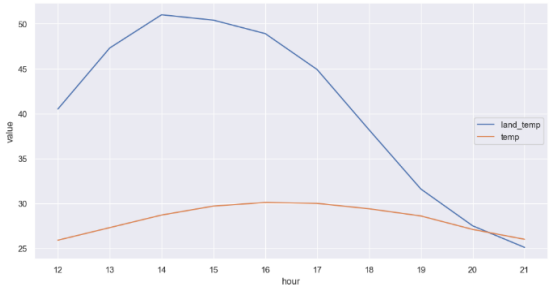
3. 여름 모델링

각 계절별 개별 모델링에 대한 성능을 파악할 기준을 만들기 위해 24개의 lag를 사용하여 계절 구분 없이 전체 데이터를 통해 만든 단일 모델을 베이스 모델로 두고 test MAE(Training data에서의)를 비교해 주었다. 여름 또한 봄과 마찬가지로 강수와 습도의 영향을 크게 받는다고 생각해 온도, 습도, 강수, 일사량 같은 지표들에 이슬점 온도까지 추가해 상호작용항을 생성해주었다. 추가적으로 기온이나 강수와 같은 변수들이 이전 시간대 대비 얼마나 급격하게 증가하고 감소했는지를 반영할 수 있도록 증감률을 고려한 pct_temp와 같은 파생변수를 만들어 사용했다. 앞서 제시한 봄 모델링과 마찬가지로 Area별 분포의 차이가 어느정도 존재하는 것으로 판단하여 Area별로 각각 LightGBM을 통한 모델링을 한 후에 앙상블을 통해 단일 모델을 만들어 주었다.

하지만 베이스 모델의 test MAE를 뛰어넘는 개선된 모델을 찾기가 어려웠다. 오히려 여름 모델링은 여름 데이터에 국한되지 않고 다른 계절의 데이터를 함께 사용한 모델이 성능이 더 좋게 나오는 것을 확인할 수 있었는데, 이는 다른 계절에서 찾아볼 수 없었던 여러 패턴들에서 기인한 결과라고 생각했다. 데이터 중 잔차가 큰 부분을 그래프로 그려 확인해보니 다소 납득하기 힘든

패턴들이 다수 있는 것을 확인할 수 있었다. <그림3>과 같이 갑자기 다른 기상지표의 별다른 변화가 없는데도 불구하고 지면온도가 급감하는 경우나 <그림4>와 같이 오후 2시인데도 불구하고 지면온도가 급감하는 추세를 보이는 지점이 존재하는데, 이러한 패턴의 경우에는 일반적인 여름의 패턴에서 벗어나 있기 때문에 오히려 여름 데이터에 과적합 하지 않고 전체 데이터를 사용하여 다소 general한 모델을 구축하는 것이 낫다는 결론을 내렸다. 따라서 최종적으로 기존에 만들었던 전체 데이터를 이용한 lag24 단일 CatBoost모델을 최종 모형으로 확정했다.





<그림3>,<그림4>-모델링으로 예측하기 힘든 패턴들 (여름)

4. 가을 모델링

가을의 경우, 일교차가 큰 날씨와 함께 8월의 강우 등 복잡한 날씨 변화를 정확히 반영하기 위해 기온, 이슬점온도, 상대습도, 풍속, 강우, 일사량, 일조 시간 변수에 대해 1시간부터 24시간 전까지 모든 시간대의 기상 변수를 고려했다. 8월은 9,10월과 달리 비가 내리는 경우가 많아 일반적인 패턴과 다른 기상현상이 발생할 수 있다는 점에서 특정 시간대만 고려하는 경우 노이즈를 학습하게 될 수 있으므로 24시간 동안 이전시간의 기상상황을 반영하고자 했다. 또한, 9,10월의 경우 강한 햇빛으로 지면이 가열되어 지면온도의 상승으로 이어지는 경우가 많음을 반영하기 위해, 3시간동안의 일사량 누적합을 구하여 변수로 활용했으며, 이와 함께 상대습도, 기온, 풍속과의 상호작용도 변수로 추가했다.

한편, 가을의 8월은 9,10월에 비해 비가 내리는 경우가 많아 비교적 상대습도 및 비와 관련된 변수의 설명력이 필요하기에, 상대습도와 비, 이슬점 온도, 풍속의 상호작용항을 통해 습도로 인한 기온의 변동 폭 축소와 강우로 인한 기온 저하를 설명하고자 했다. 또한, 기온, 일사량과 일조시간을 한번에 고려한 3차원 상호작용 변수와 상대습도, 기온, 풍속의 3차원 상호작용, 기온, 상대습도, 일사량의 3차원 상호작용항을 통해 가을철의 높은 일사량과 강우, 습도가 지면온도에 미치는 영향력을 효과적으로 설명하는 변수를 생성할 수 있었다

다음으로, 위에서 설명한 상호작용항 중 상대습도와 비, 상대습도와 이슬점 온도의 상호작용항의 경우 비가 지속적으로 내려 상대습도가 증가하고, 대기가 건조하지 않은지를 연속적으로 반영해야 함을 고려했다. 따라서, 해당 상호작용항에 대해서는 최근의 정보가 지면온도 예측에 큰 영향을 미칠 것이라는 가정하에, 최근 3시간의 값과 24시간 전의 기상상황은 현재와 유사한 형태임을 확인하여, 24시간 전의 값을 lag로 사용했다.

마지막으로, 앞서 설명한 전처리 방법 중 가을 모델링에는 2-3)의 month, day 변환을 적용했으며 2-7) 기타 파생변수에서 소개한 기온과 이슬점 온도의 차이를 사용했다. 최종적으로 Bayesian Optimization 방법을 이용해 area별 LightGBM 모델의 최적 hyperparameter를 찾아 주었다.

5. 겨울 모델링

건조한 날씨일수록 일간 지면온도 차가 커지기 때문에 겨울은 기온, 일사량뿐만 아니라 이슬점 온도, 습도에도 많은 영향을 받는다. 따라서 위 변수들에 대해 이전 시간대(1, 2, 3, 6, 12, 24시간) 대비 증감률 변수를 사용했다. 위 변수들과 강수량, 적설량 변수에 대해 3시간 가중평균을 구하여 변수로 활용했다. 또한, 기온, 습도, 풍속, 일사량과 같은 변수들의 상호작용 항도 추가했다. 이를 통해 겨울철에 낮은 기온과 높은 습도 조건 하에서 지면에 서리가 형성되거나 눈이 녹지 않아 지면온도가 추가로 하락하는 상황 등을 모델이 감지할 수 있도록 해 주었다.

마지막으로 앞서 설명한 전처리 방법들 중 효과적이었던 것들을 추가해주었는데, 비와 눈과 관련된 2-7)의 after_snow, after_rain, 누적 강수량, 누적 적설량 변수와 눈이 오는 경우 보온효과가 일어나는 현상을 반영해주기 위해 눈이 온 시간대에 대해 2-5) 전처리 과정을 반영해주었다. 2-2)에서 언급한 Wavelet 변환 또한 효과적이었기에 채택하여 모델링에 사용했다. 최종적으로 Bayesian Optimization (Tree-structured Parzen Estimators) 방법을 이용해 area별 CatBoost 모델의 최적 hyperparameter를 찾아 주었다.

6. 최종 모델링 결과

	봄	여름	가을	겨울	전체
모델	CatBoost+Ensemble	Full data + Single CatBoost model	LightGBM + Ensemble	CatBoost+Ensemble	
검증 MAE	1.633	1.845	1.617	1.654	1.687

IV. 활용 방안 및 기대효과

본 프로젝트에서는 다양한 통계적 방법을 적용하여 새로운 지역의 지면온도를 정확하게 예측할 수 있는 모델을 제시했다. 탐색적 데이터 분석을 통해 지역에 따라 각 기상 변수의 분포가 차이가 있음을 발견했고 이 차이를 통제하기 위해, 각 지역 데이터를 이용해 base model을 만든 다음 앙상블하는 방식을 채택했다. 시간에 따른 주기성을 반영하기 위하여 푸리에 변환을 적용했으며 데이터 자체의 노이즈를 줄여주는 Wavelet 변환 또한 효과적으로 작용했다. 또한 상호작용항이나 파생변수 같은 각 계절에 특화된 변수들을 만들어 계절의 특성을 적절하게 반영해 주었다. 개별 계절에 대한 모델링이 전반적으로 성능이 좋은 모습을 보여주었지만 여름과 같이 데이터의 예측하기 어려운 특이한 패턴들이 발생하는 경우에는 계절 하나에 국한되어 모델링을 하기 보다는 전체 데이터를 통한 일반적인 모형이 성능을 향상시킬 수 있는 점도 확인할 수 있었다. 위와 같은 모델링 과정을 통해서 기존에 보유하고 있는 지역 데이터를 사용한 모델을 사용하여 동일한 지역, 더 나아가 다른 새로운 지역의 지면온도 예측에 활용이 가능할 것이라 기대한다. 추가적으로 결측값 처리를 위한 GAN 모델을 활용한다면 시스템 오류 등으로 인한 결측값을 효과적으로 해결하여 더욱 안정적인 지면온도 예측이 가능할 것이라 생각한다.