**과제 1. 생활안전(지표면 온도 예측 AI 모형 구축)**

**팀 명:Data trip**

**Contents**

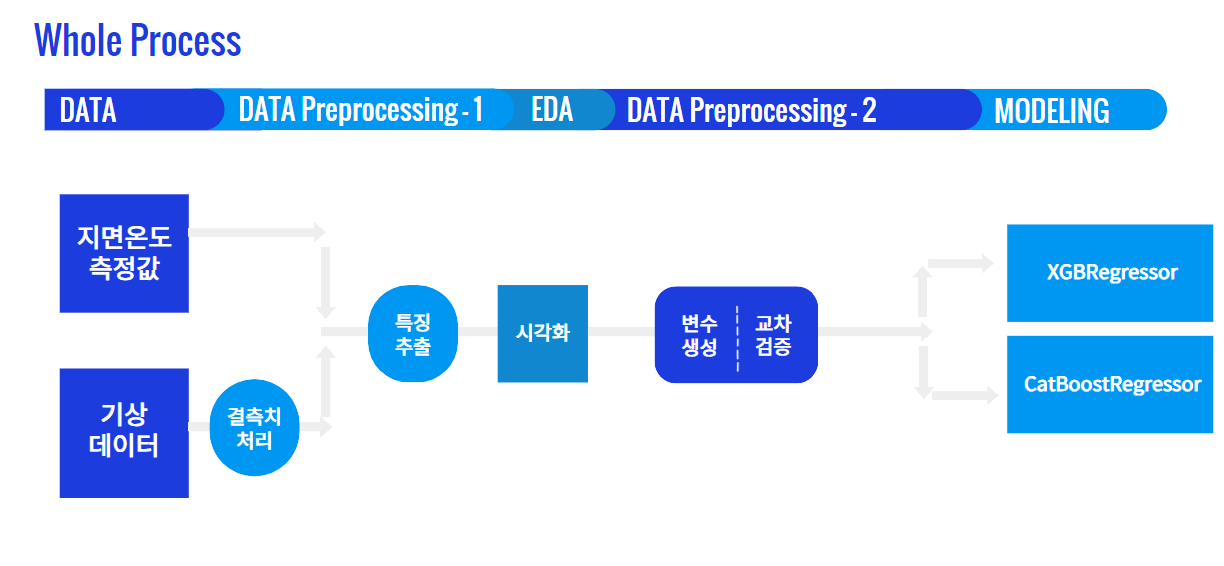
* **1 Introduction**
* **2 Load and check data**
  + 2.1 Load data
  + 2.2 Description
  + 2.3 Outlier detection
  + 2.4 Check for null and missing values
* **3. Data preprocessing & Feature engineering**
  + 3.1 Data preprocessing
  + 3.2 Feature engineering
* **4 Feature analysis**
  + 4.1 Numerical values
  + 4.2 Categorical value
  + 4.3 Correlation
* **5 Modeling**
  + 5.1 Simple modeling
    - 5.1.1 Cross validate models
    - 5.1.2 Based regressor (CatBoostRegressor)
    - 5.1.3 Feature importance of the tree based regressor
  + 5.2 Ensemble modeling
    - 5.2.1 Combining models (CatBoostRegressor, XGBRegressor)
  + 5.3 Prediction
    - 5.3.1 Predict and Submit results
* **6. Conclusion**
* **7. References**

**1 Introduction**

국내 10개 지점의 5년간 11개 기상 데이터를 전부 혹은 일부 사용하여 계절별 1개의 통합 산출 모델을 개발하고, 예측값을 도출한다. 평가지표는 예측값과 실제값의 차이의 절댓값을 평균으로한 MAE 지표를 사용한다.

현재 기상청은 기온관측 지점에 비해 지면온도의 관측 지점 수가 적다. 만약, 기상자료를 활용한 지면온도 산출 기술이 개발된다면, 관측 공백 지역의 지면 온도값을 추정할 수 있을 것이다. 이는 실생활에 밀접한 여름철 폭염 및 겨울철 한파 대응에 도움이 될 것으로 기대된다.

사용데이터는 기온과 이슬점온도, 상대습도 등의 기상데이터를 사용하였으며, 특성 공학을 이용하여 변수를 추가하였다. 이후 앙상블 모델로 예측하고자 한다.

**Whole Process**

**2 Load and check data**

**2.1 Load data**

제공된 suface\_tp\_train, surface\_tp\_test 데이터를 사용하였다.

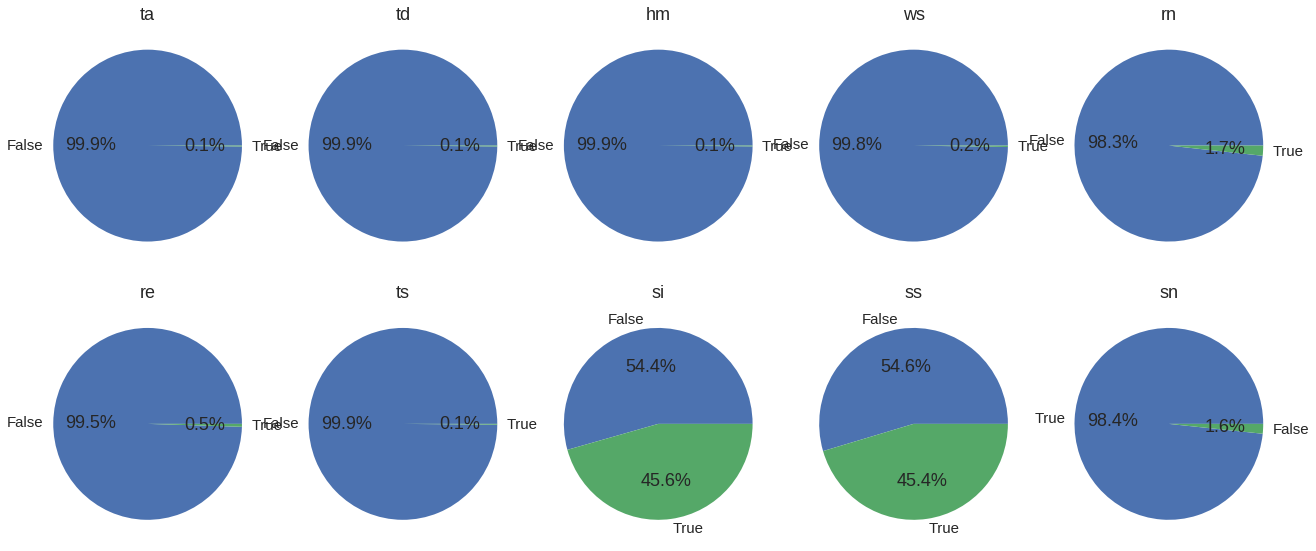
**2.2 Description**

`sufrface\_tp\_train.stn`의 경우 10개 지점에서 측정이 되었으며 -99.00 또는 -99.9의 값을 가지고 있는 변수들이 존재했다. 즉, 결측치가 존재하는 것을 확인할 수 있었다.

**2.3 Outlier detection**

`surface\_tp\_train.ta`(ta) 변수와 `surface\_tp\_train.ts’(ts) 변수 간 산점도를 확인했을 때, 이상치가 존재하는 것을 확인할 수 있었다.

**2.4 check for null and missing values**



-99.9, -99의 값을 모두 결측치라고 간주하고, 총 결측 비율 확인했다. si, sn 의 라벨이 45% 이상의 높은 결측치를 가진 것을 확인했다.

**3 Data Preprocessing & Feature engineering**

**3.1 Data preprocessing**

* train set과 test set 모두 변수명을 올바르게 교정하였다. 예를 들어 `surface\_tp\_train\_tp` -> `tp` 로 변경하였다.
* mmddhh 변수에서 month, day, hour 변수를 추출하였다.
* `ww` 변수에서 희소한 레이블들은 하나로 묶어주었다.
* `ta` 변수와 `ts`의 값이 -80 미만인 값들은 이상치로 판단하여 제거하였다.

|  |  |
| --- | --- |

왼쪽은 이상치를 제거하기 전이며, 오른쪽은 이상치를 제거한 후의 산점도이다.

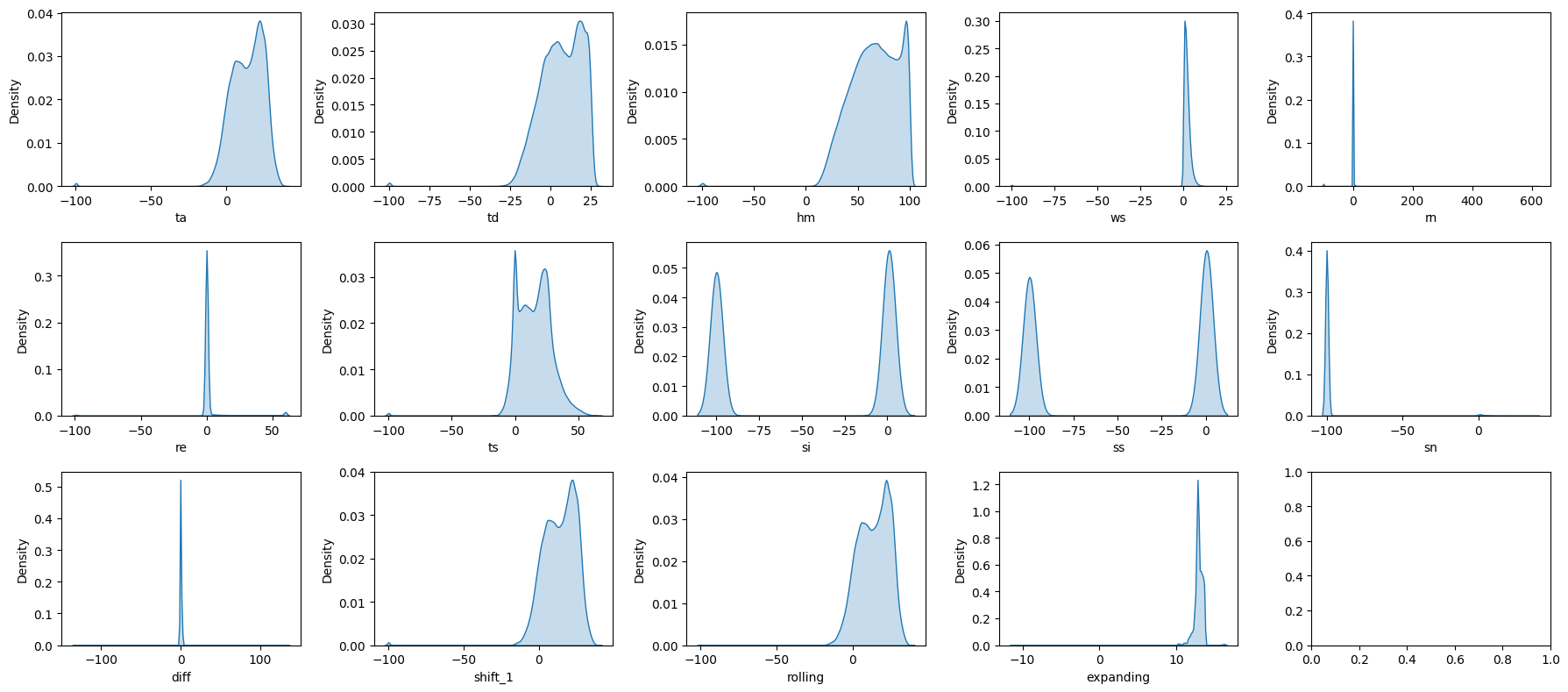
* Boosting 모델은 -99.9 또는 -99 값들을 수치로 계산하기 때문에 NaN으로 변환하였다.
* 범주형 변수는 LabelEncoder로 수치형 변수로 변환하였다.

**3.2 Feature engineering**

* 본 시계열 데이터의 변화율을 확인하고자 1차 차분 변수를 생성하였고 이는 현재 시점의 값과 이전 시점의 값과의 차이를 계산한 변수이다.
* 현 시점과 다음 시점 간의 관계를 파악하기 위해 시차이동 변수를 생성하였다.
* 이동평균을 계산하고자 이동통계량(rolling)과 현재 시점까지 확장되는 평균(expanding) 변수를 각각 생성하였다.

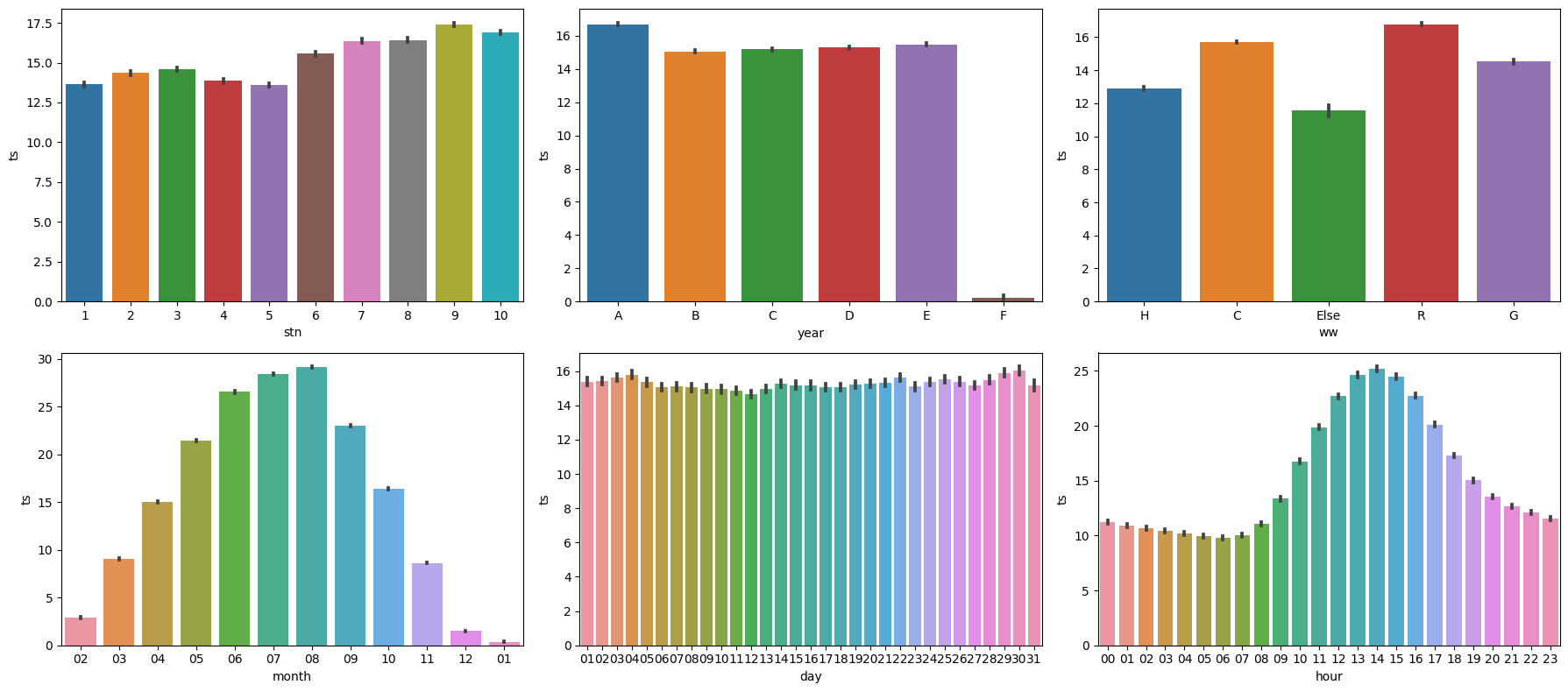
**4. Feature analysis**

**4.1 Numerical values**



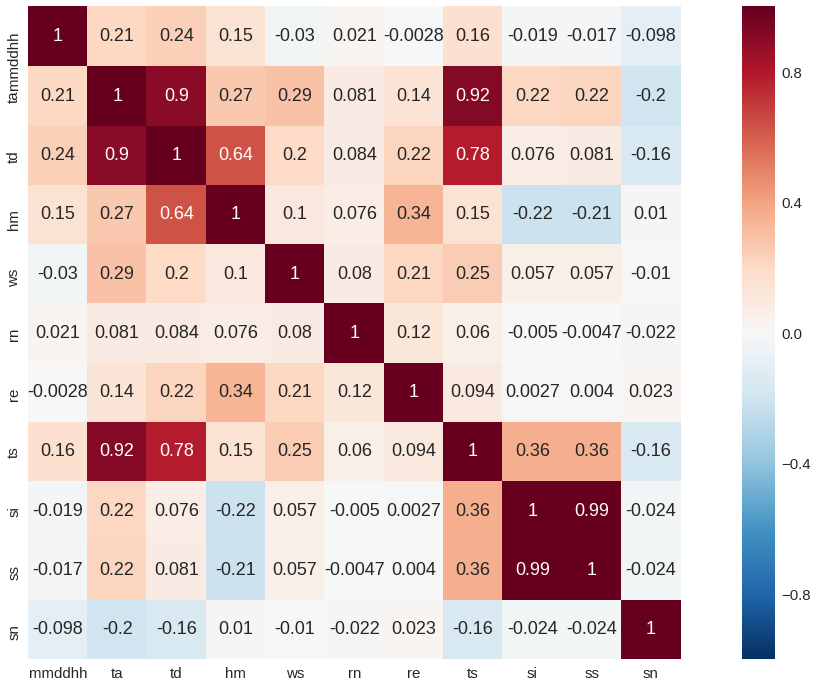
각 데이터의 커널밀도추정(KDE)를 활용하여 각 변수의 분포를 확인하였다. 대부분의 변수에 이상치(-99.9 또는 -99)가 존재하였으며, 오른쪽으로 치우쳐져 있다. (data skewed)

**4.2 Categorical value**



지면온도를 막대 차트로 확인해본 결과, `stn` 지역 번호에 따라서 상이하며, `day`에서 총 값은 비슷한 반면, `month`과 `hour`에서 시간에 따라서 큰 차이를 보이는 것을 확인하였다.

**4.3 Correlation**



* 주어진 변수의 상관분석을 통해 변수들 중에서 지면온도(ts)와의 관련성이 높은 변수들을 선택하고, 필요한 경우 추가적인 특성 공학을 진행하였다.
* 지면온도(ts)와 양의 선형관계를 가지는 변수는 순서대로 `ta`,`td`,`si`,`ss`순이다.

**5. Modeling**

**5.1 Simple modeling**

**5.1.1 Cross validation**

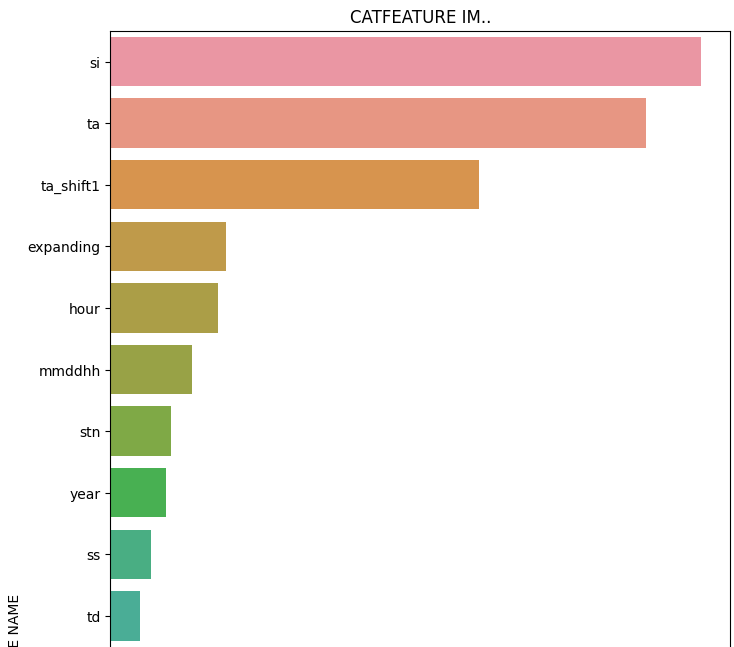
10 folds 교차검증을 사용하였으며, 평가지표는 MAE(Mean Absolute Error)를 사용하였다. 아래 그림은 10 folds 교차검증에 대한 그림이다.

**5.1.2 Based regressor (CatBoostRegressor)**

XGBRegressor, LightGBM, CatBoostRegressor, ExtratreeRegressor 네가지 회귀모델을 사용하였고, 가장 낮은 MAE값을 보여준 CatBoostRegressor을 베이스라인 모델로 선정하였다.

**5.1.3 Feature importance models of the tree based regressors**

계절 중 하나인 봄을 CatBoostRegressor로 Feature importance를 확인해보았다.



- 변수 중요도는 `si`, `ta`, `ta\_shift\_1` 순이며, `si`변수에서의 결측치는 중요한 정보중 하나인 것을 알 수 있었다.

**5.2 Ensemble modeling**

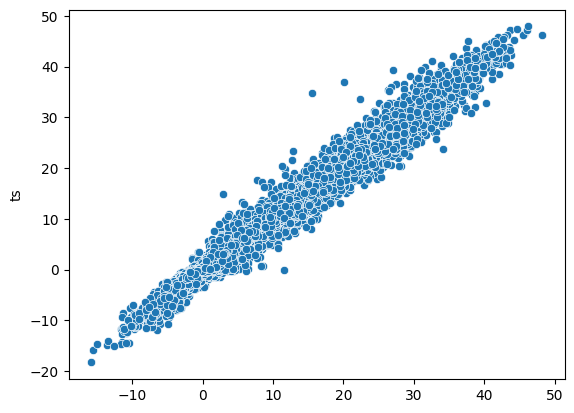
5.2.1 Combining models (CatBoostRegressor, XGBRegressor)

| #Ensemble modeling  from sklearn.ensemble import VotingRegressor  vot\_model = VotingRegressor(estimators=[('XGBoost',XGBRegressor(random\_state=42, tree\_method='gpu\_hist')),('CatBoost',CatBoostRegressor(random\_state=42, silent=True,task\_type='GPU'))],  weights=[0.2, 0.8]) |
| --- |

XGBRegressor과 CatBoostRegressor을 앙상블하였으며, 가중치는 각각 0.2, 0.8로 두었다.

**5.3 Prediction**

**5.3.1 Predict and Submit results**

최종적으로 각 계절들을 VotingRegressor 모델을 사용해서 예측을 하였으며, 얼마나 비슷한지 실제값과 예측값의 산점도를 확인하였다. 아래 산점도는 계절이 봄일 때, 지면온도의 실제값과 예측값을 확인한 그림이다.

- 이후 모든 계절들을 위와 같은 Ensemble 모델로 모두 예측하였다.