We collect 3 datasets from 3 different monitoring stations (NK,ST,WL). To be specific, NK has the sample size of 4187 and ST of 1811 and WL of 1256.

All these three datasets have the same 18 variables that monitored and collected by the stations. The 18 variables can be separated into 2 parts. The first part is our target variables: *ultrafine particles numbers per cm3(number)* and *the dominant particle size when the measurement was taking place(size).*

The rest 16 variables are environment related and date related variables such as temperature, humidity, season, weekdays etc.

Due to the small sample size of our datasets, we consider not to apply deep neural networks for our later model training part.

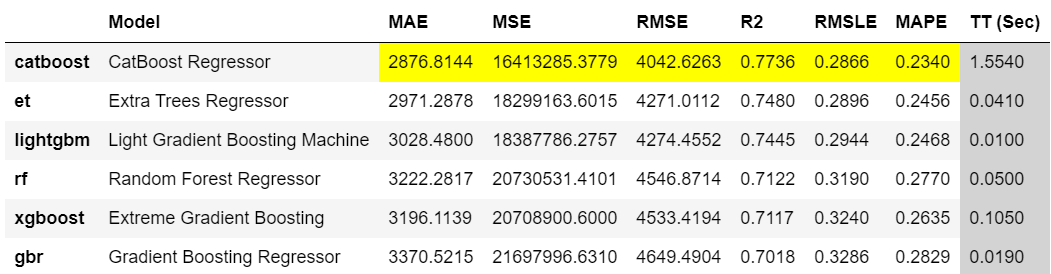
We perform typical machine learning models (regression models) to see

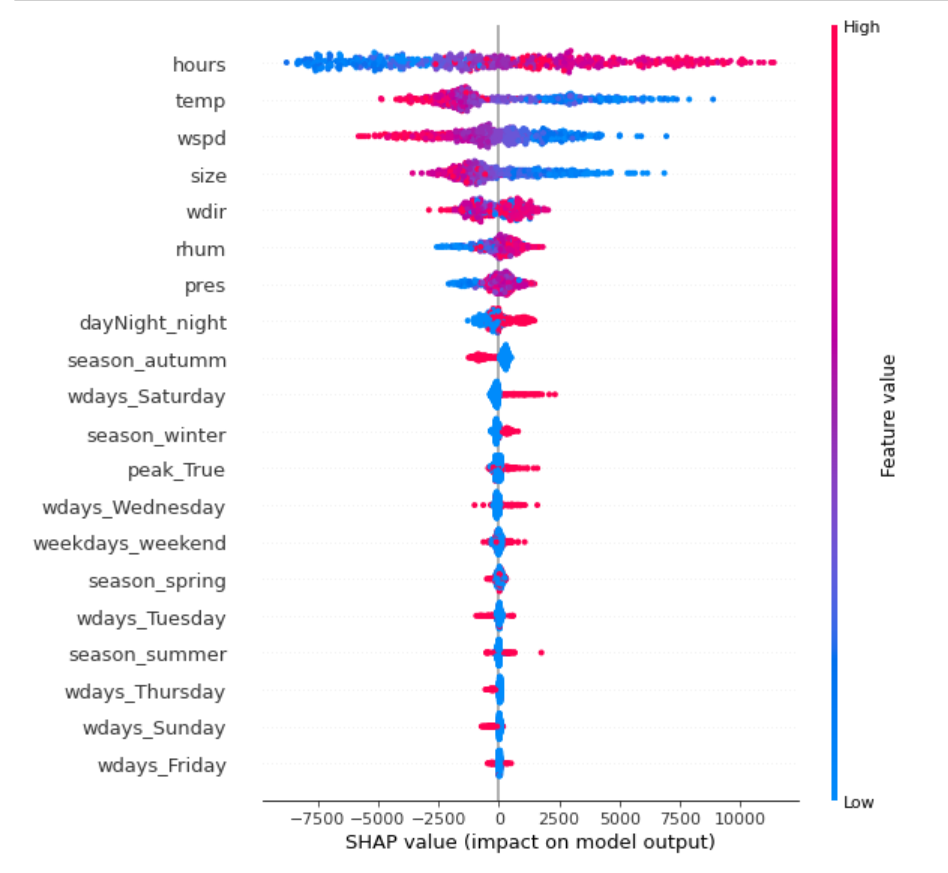
1. the possibility of predicting *number* and *size*
2. the impacts that other variables on our target variables.

We apply regular steps including data cleaning, data preprocessing before model training, and we train different models using 10-fold cross-validation with default hyperparameters to compare the baseline performance of these models. We use shap values to get a relative straight ideas of the impact of those 16 variables on our model output.

* *Numbers* as target variable (include size):

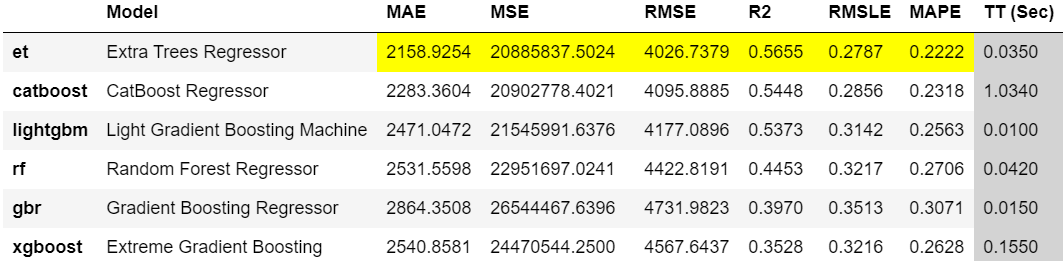
ST

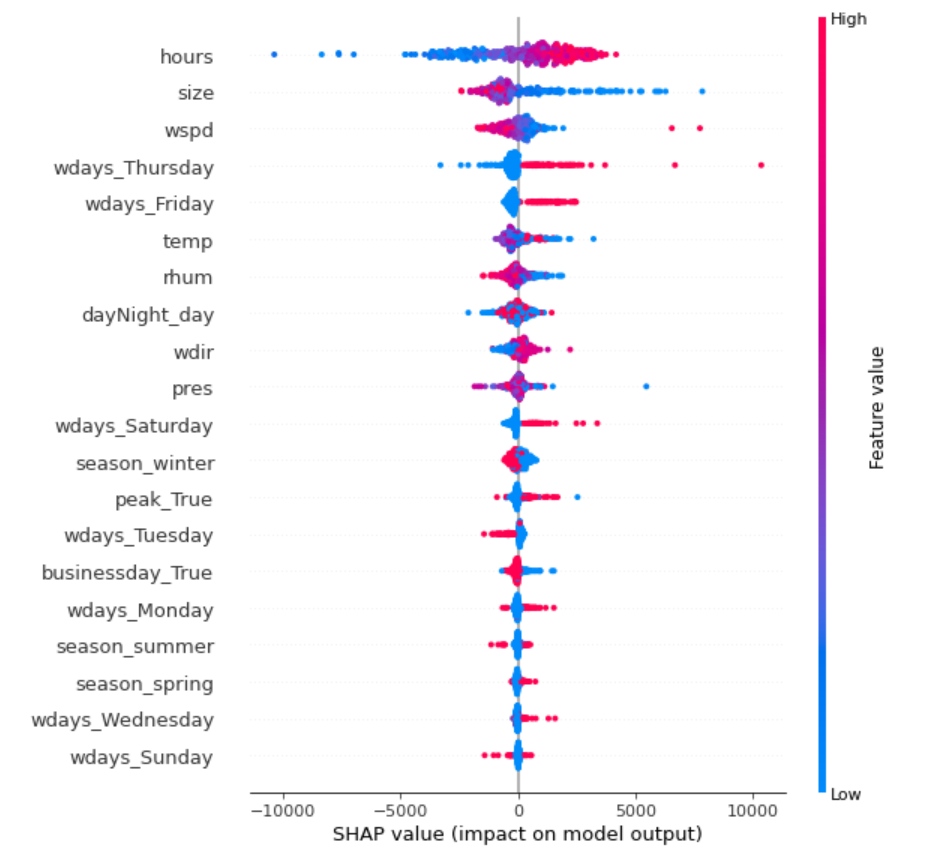




Top 3 variables affect *numbers: hours, temp, wspd.*

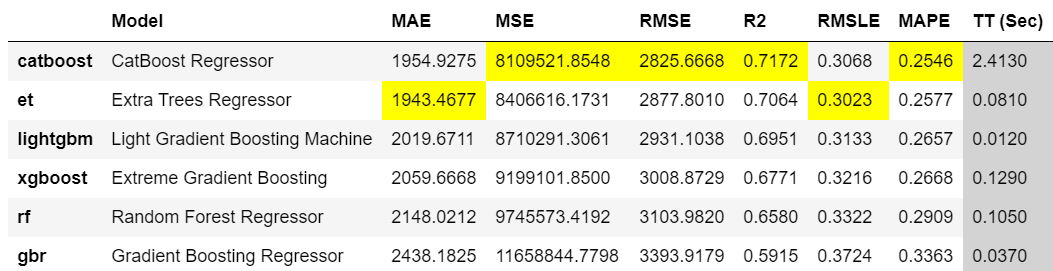
WL

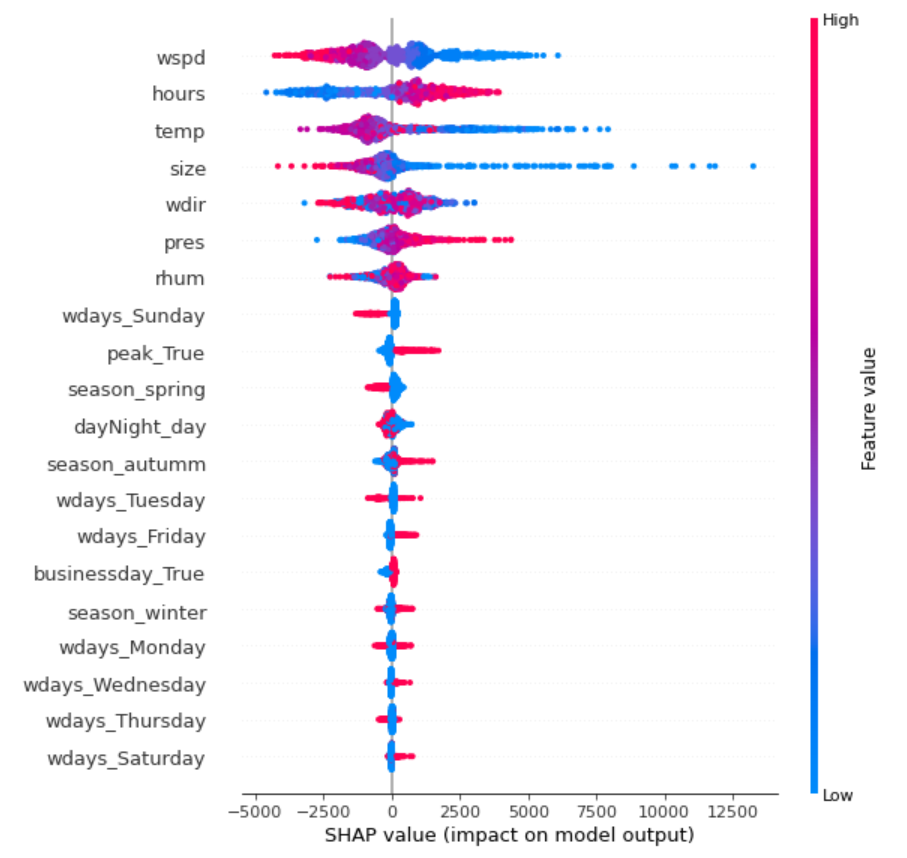




Top 3 variables affect *numbers: hours, size, wspd.*

NK

****

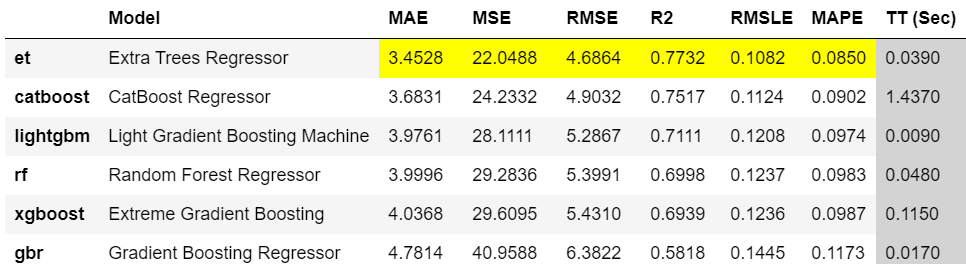
****

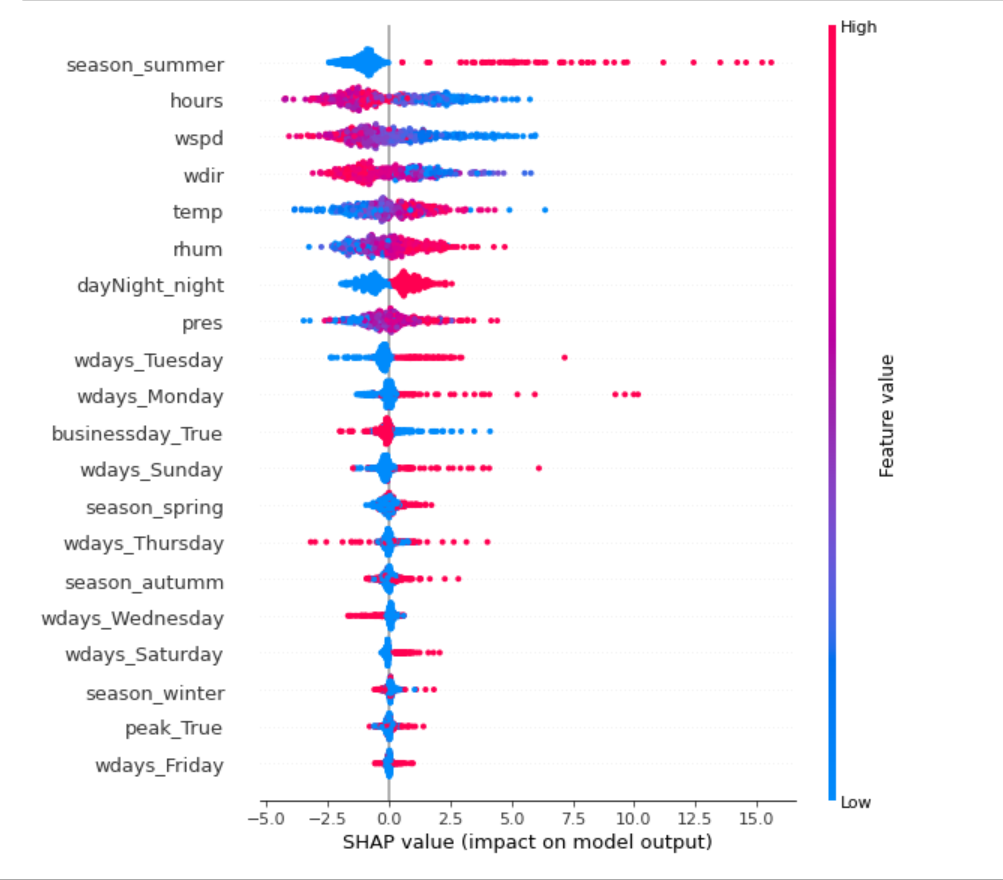
Top 3 variables affect *numbers: wspd, hours, temp.*

We can see from the above pics that both NK and ST dataset got a relative higher R^2 above 0.7 while WL only got R^2 above 0.5. We can also expect that tree based models especially catboost regressor and extra trees regressors have a better base performance than other models. At last, *hours* and *windspeed* might be the most important factors that affect the ultrafine particle numbers.

* *Size* as target variables (exclude numbers):

ST

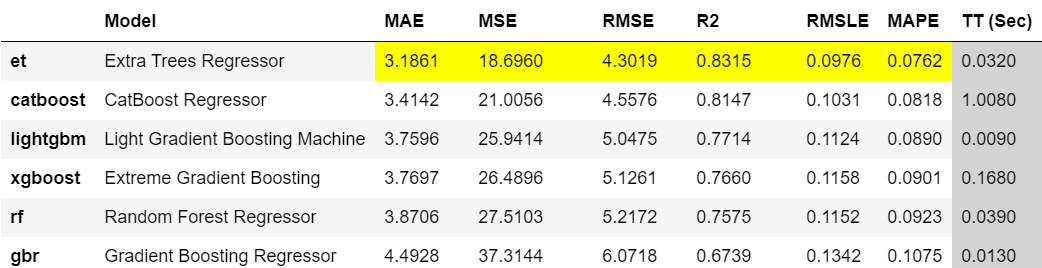
****

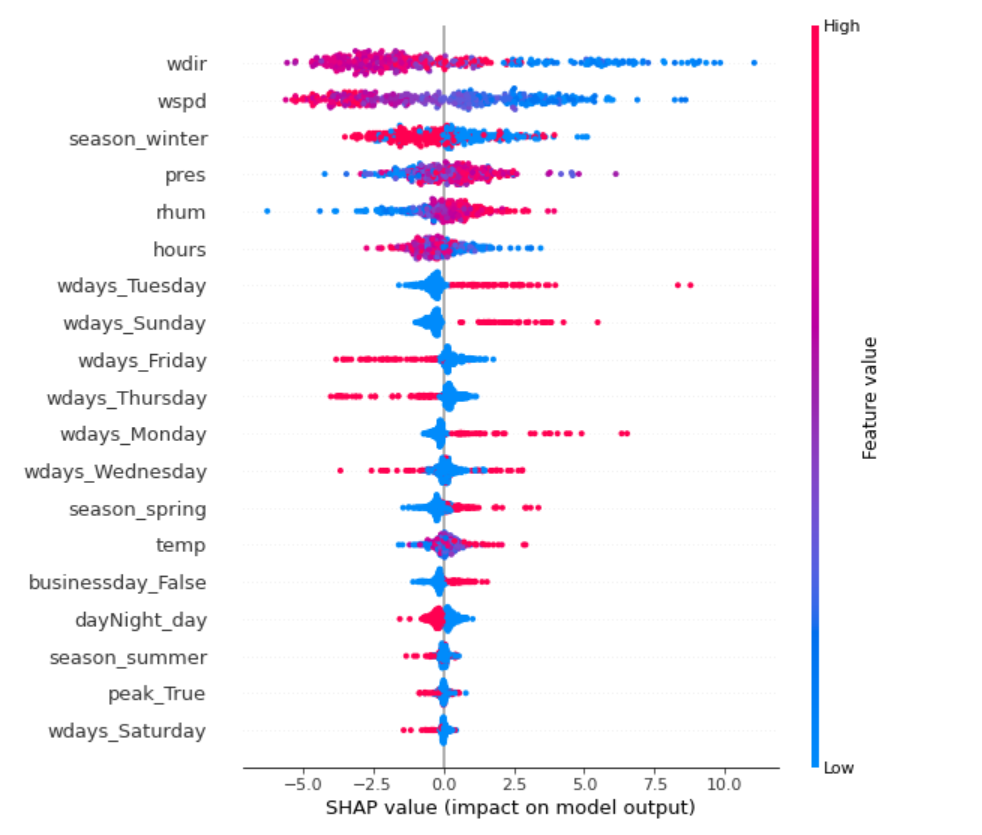
****

Top 3

Top 3 variables affect *numbers: season, hours, wspd.*

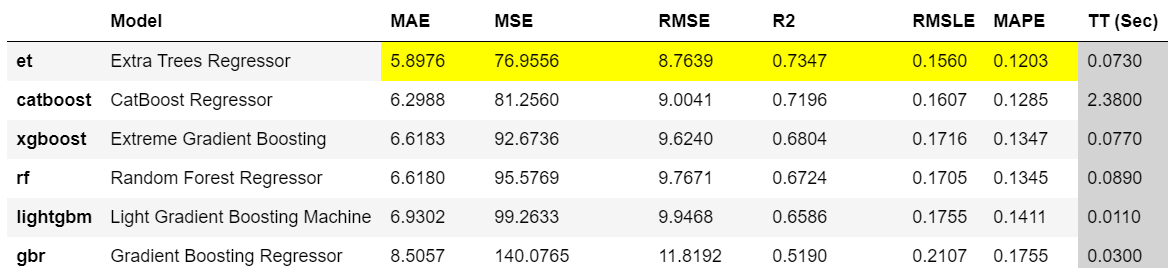
WL

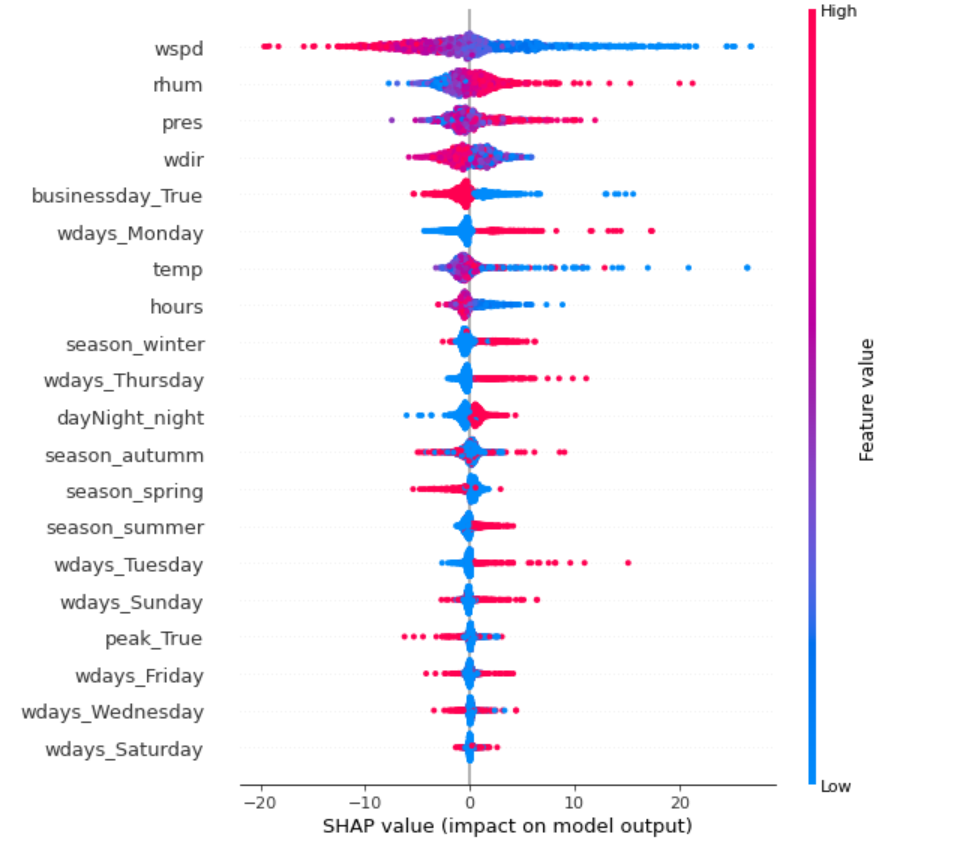
****

****

Top 3 variables affect *numbers：wdir, wspd, season.*

NK

****



Top 3 variables affect *numbers: wspd, rhum, pres.*

From above pics, we notice that if we treat size as our target variable and exclude number when we train our model, the models performance (R^2) get better than our first scenario (all above 0.7, R^2 of WL even up to 0.8). The top 3 factors that affect size vary differently and do not indicate consistency with the first scenario but we can still see that *windspeed* still plays an important role if we want to predict *size* or *numbers.*