1. Introduction:

Problem Statement:

To build a machine learning model to forecast the soil moisture level given a time series data of daily soil moisture measurements at a specific location from July 2022 to March 2023.

Soil moisture is the measure of water content in the soil that tells about how much life can sustain with that soil.

Knowing about soil moisture can help humans in many ways like:

Farmers can make precision for irrigation, thus can help in cutting down water wastage. Soil acts as the main content that decides vegetation in an area.

Analyzing the flood and drought conditions can help in early warnings and taking precautions.

Biologists and environmentalists can use the soil moisture level to analyze the type species living in the habitat and what impact it makes on them due to change in the soil moisture level.

2. Exploratory Data Analysis:

Seasonality:

It refers to the repeating pattern or trend over time of a parameter.

Our target variable (sm) shows seasonality between -14 and 11 with respect to time. This pattern places important role in training the model and improving the accuracy.

Stationarity:

It refers to the statistical properties like mean, variance etc of the variable to be stationary with respect to time in a time series data.

The variable changing with time will not add any significant value to the learning model.

For our data, stationary is tested using Dickey-Fuller Test, it is found that pm1, pm2, and pm3 attributes have decreasing stationarity with time. It is necessary to remove the stationarity in them by methods like differencing and log transformation.

Trend:

It is another measure to say whether the data is stationary. Identifying the trend is an important step in analyzing and modeling the data. Trends can provide insights into underlying patterns and can be used to make predictions about future values.

There is a negative trend for attributes pm1, pm2, and pm3 which confirms that data is not stationary.

Correlation between attributes:

Heat maps are the best way to show the relation between the attributes of the data.

It was found that pm2 and pm3 have a strong, positive correlation with the target variable. st, am, and pres have negative correlation with target.

3. Data Preprocessing:

Null Values:

Null values can lead to bias analysis and affect the accuracy of the model. Hence it is important to remove null values or replace them with appropriate values. The given data does not contain any null values and we are good to go.

Outliers:

They can lead to inaccurate results and overfitting of the data. Removing them will make data scaled. Outliers of the given data are visualized using box plot and removed by interquartile range method.

Scaling:

In a given data it could so happen that larger values of the attribute may dominate over smaller values. Especially for the methods that use averages or mean values. Hence, it is important to scale the data.

The Standard Scalar method is being used to scale our data.

3. Model Selection:

For time-series forecasting, models like ARIMA, SARIMA are the best known models. Also there are many regressors right from Light regressor through XGboost, Light GB, random forest, Decision tree, Catboost etc can also be used. Moving to DLL we have CNN, RNN, LSTM and transformers models for time series analysis.

Since the data is too small for DL models like LSTM and transformers, the models cannot capture enough information for making predictions.

Also as our data is not stationary, and the variable showing high correlation with target variable show decreasing trend. For models ARIMA and SARIMA the data needs to be stationary and show no trend. Hence, we cannot use these models for prediction.

We are left out with regression techniques and we went for ensembling them together.

4. Model Evaluation:

The two data files given have different times. Hence to combine them we encoded the time to unix timestamp. We split the data into train, validation and test set.

We separately trained Linear Regressor, Decision Tree, Random forest, XG boost and Light BG models and took the mean of their predictions as ensembled prediction and final predictions.

5. Conclusion:

It was fun to work on real time data. The challenge was the size of the data and separated with different attributes. Our model has not shown expected results since preprocessing and feature engineering was not done satisfactorily. But we had a lot to learn and a lot to explore on how we can train small data with and without merging them. We would like to thank the hosting team for letting us take part in the contest.