Zelestra X AWS ML Ascend Challenge - 2nd Edition

Programming Language: Python

Libraries: numpy, pandas, matplotlib, seaborn, scikit-learn

Algorithms: Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, Adaboost Regressor, XG Boost Regressor

Tools: Jupyter notebook, VS Code

Folder Structure:

```
root/
│
├── artifacts/
│         ├── data.csv
│         ├── test.csv
│         ├── train.csv
│         ├── preprocessor.pkl
│         └── model.pkl
│
├── src/
│         ├── exception.py
│         ├── util.py
│         └── components/
│                   ├── data_transformer.py
│                   ├── data_ingestion.py
│                   └── model_trainer.py
│
├── sample.csv
```

```
 ├── test.csv
 ├── train.csv
 ├── zelestra.ipynb
└── requirements.txt
```

Preprocessing Steps

1. I have used .describe() method of pandas and found that some numerical features are having data type as object, so first of all changed the data type of those columns to number.
2. Then I checked for null values and found that for categorical features error_code and installation_type there are almost 25% of the data is missing and also observed using bar graph that other 3 values in these features are equally distributed so I imputed these missing values with new value ‘NA’ creating 4 different values in these categorical features.
3. For numerical features I have observed that approx. 5% values are missing this seems to be good amount of missing data so I have used KNNImputer to impute missing data for these numerical features.
4. By finding correlation I observed that temperature and module_temperature are highly correlated so I dropped module_temperature column. And also observed that wind_speed and pressure are very weakly correlated with efficiency so I dropped these 2 columns.
5. Then for categorical Features I have used one hot encoding.

Training Model:

I have used different Machine Learning algorithms with different hyperparameters and evaluated the result of every model and out of those chose the model with best score for our prediction. Machine Learning algorithms that I tried are Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, XG Boost Regressor, Catboosting Regressor, Adaboost Regressor.