**Power Prediction for weather forecast data**

**Datasets:** Following three files were given:

* **File1:  power\_actual**

1. This file contains the solar generation of a certain plant from October 1st, 2017 to September 30th, 2019.

2. You'll find the following columns: 'power', 'gti' and 'ghi'. Power is the actual power generated while GHI (Global Horizontal Irradiance) and GTI (Global Tilt Irradiance) are the parameters relevant to the that define the radiation received from the sun.

* **File2: weather\_actuals**

1. This file contains the weather data of the same plant from October 1st, 2017 to September 30th, 2019.

2. The columns' names are self-explanatory.

* **File3: weather\_forecast**

1. This file contains the weather data from October 1st, 2019 to October 27th, 2019.

**Problem Statement**: You need to predict the generation of power of the given plan in this duration: October 1st, 2019 to October 27th, 2019.

**STEPS FOR CREATING MODEL:**

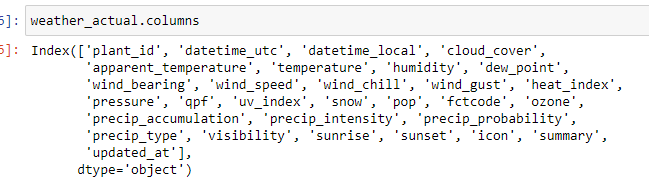
1. Reading the data.
2. Removing the features whose data is missing significantly in weather actual data.
3. Replacing the incorrect value (-9999) present in the weather actual data with meaningful values.
4. Detecting and handling outliers of weather actual data.
5. Checking the power data.
6. Comparing the dates in power and weather actual data.
7. Merging the power column from power data to weather actual data, to perform EDA
8. Feature Engineering
9. Choosing right Machine Learning model.
10. Predicting the power for the weather forecast data.

**Details:**

1. Reading the data:

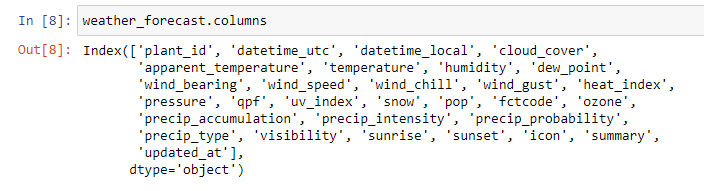
Weather data: Shape of the dataset is (13619, 30)

Columns:



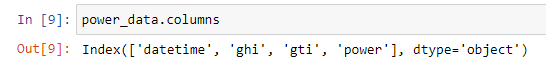
Weather forecast data:

Columns:



Power data: shape of the data (70080,4)

Columns:



1. Removing the features whose data is missing significantly in weather actual data.

There were 8 features whose data was missing significantly. Around 72% of data was missing. So removed those features from the dataset.

1. Replacing the incorrect value (-9999) present in the weather actual data with meaningful values:

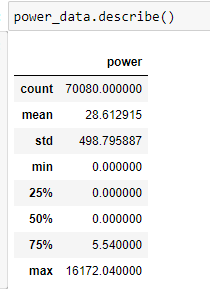
First replaced all the -9999 value with NaN values. So that these values will be treated as missing values by the python. Interpolated these missing values with linear method. Among all the method present in the interpolation (pandas api) linear method did not change the data distribution.

1. Detecting and handling outliers of weather actual data:

From the box plot, we can infer that there were some outliers present in the some features. Capped the value of all the outliers by (Q3 + 1.5 IQR).

Capping was a good solution as the given data is a weather data. It won’t change significantly for some period of time.

1. Checking the power data:
   * Total number of missing values: Feature ghi and gti had 50% of the data missing. So dropped these columns from the power data. From the power description we can infer that there are few high outliers present in the power data.
   * Description: The power mean is 28 and the standard deviation is 498. Which means that there are few but high values (outliers) present in the power dataset.



* The power in power data is recorded in every 15 minutes, whereas in weather actual and forecast data the feature values are recorded hourly. To merge the power values with weather actual values, sum the power value hourly by using group by and sum function in python.

1. Comparing the dates in power and weather actual data: There are few missing dates in the weather actual data.

* From 01-10-2017 to 30-09-2019 there are 730 days. So the total column we should expect in the dataset: 730 \* 24 = 17,520.
* In power data we have 17520 row. That means we have no missing date in power data.
* But in weather data there are 13619 rows which means there are some missing dates in it.
* Total number of missing rows : 17520 - 13619 = 3901.
* There are two possibility for these missing rows: 1) Either whole day data is missing 2) Partial day data is missing.
* For whole day data missing. Compared the dates present in the power data and weather data. Then made the list of missing dates.
* For Partial day data missing, Checked the value counts of the date. If the value count was not 24 then added those dates to missing dates. For keeping the dates similar in both dataset, removed these missing dates from weather actual data also.
* Then drop these dates from the power data, in order to make dates present in power data and weather data same.

1. Merging the power column from power data to weather actual data, to perform EDA:

Merge the weather actual data and power data. Perform EDA and checked the correlations between features and target variable. There were nearly 0 correlations between features and target variable. Only few features were inter-correlated.

1. Feature Engineering:

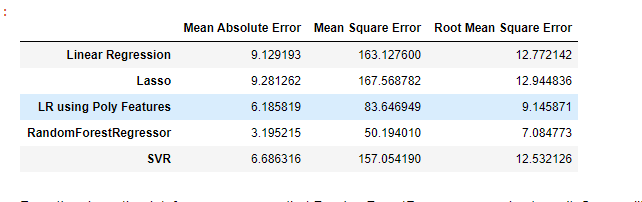
* Dropped all the irrelevant features like plant\_id, date\_time\_utc, updated\_at, apparent temperature, summary.
* Dropped the apparent temperature as it was highly correlated with temperature.
* From time stamp, extracted hour information and made an hour column. Then dropped the time stamp. According to the power distribution, the power was non-zero during the day hour. So it might help the model to recognize some pattern.
* From the sunrise and sunset features, calculated sun duration time. Then dropped sunrise and sunset features.
* Label Encoded icon column.

1. Choosing right Machine Learning model:

Checked the performance of few ML model and then selected the model which performed better. Since we needed to predict continuous variable below Machine Learning Algorithms were used:

* Linear Regression.
* Lasso.
* Linear Regression Using Polynomial Features.
* Random Forest Regressor
* SVM Regressor

Following were the model performance:



* Clearly Random Forest Regressor performed better.

1. Predicting the power for the weather forecast data.

Since Random Forest Regressor performed better than other model, predicted power for weather forecast data using Random Forest Regressor model.