

Problem Statement :

To build a machine learning/deep learning approach to forecast the total energy demand on an hourly basis for the next 3 years based on past trends

Data :

Data have captured the estimated total energy demand from the past 12 years on an hourly basis. Now, the government of Green Energy is looking for a data scientist to understand the data and forecast the total energy demand for the next 3 years based on past trends.

Approach :

Given problem is a part of Time series analysis. The model for the above problem has been created with below 4 steps.

1. Data Preprocessing
2. Feature Engineering
3. Model training and validation.

Data Preprocessing :

Input for the problem contains less number of columns.

	row_id	datetime	energy
0	1	2008-03-01 00:00:00	1259.985563
1	2	2008-03-01 01:00:00	1095.541500
2	3	2008-03-01 02:00:00	1056.247500
3	4	2008-03-01 03:00:00	1034.742000
4	5	2008-03-01 04:00:00	1026.334500

Since we have a datetime column, We have preprocessed that column before training the model.

Created a function to split the values into day,month,year,week, hour, minute

	row_id	datetime	energy	Year	Month	Day	Dayofweek	DayOfyear	Week	Quarter	Semester	Is_weekend	Is_weekday	Hour	Minute
0	1	2008-03-01 00:00:00	1259.985563	2008	3	1	5	61	9	1	1	1	0	0	0
1	2	2008-03-01 01:00:00	1095.541500	2008	3	1	5	61	9	1	1	1	0	1	0
2	3	2008-03-01 02:00:00	1056.247500	2008	3	1	5	61	9	1	1	1	0	2	0
3	4	2008-03-01 03:00:00	1034.742000	2008	3	1	5	61	9	1	1	1	0	3	0
4	5	2008-03-01 04:00:00	1026.334500	2008	3	1	5	61	9	1	1	1	0	4	0

Checked if null value is presented in the data or not. Found null values are present in target column.

```
row_id      0
datetime    0
energy     1900
```

In order to impute the null values from the target column, used an approach to replace the null values,

- A. First, grouped the target value based on hour.
- B. Find the mean of the target label based on hour and replaced the null values.

Since our problem statement denotes that the forecasting will be required on an hourly basis, chose this approach to impute the null values.

Feature Engineering

Concatenate the train and test data by creating a column 'train_or_test' with values train and test for the respective data.

As we have less number columns to process the data, created additional column using the 'Hour','Day','Minute'.

- a. Grouped the data frames with above mentioned data frames.
- b. Find Mean, Median and Mode of target label after grouping the column.
- c. Merged the Mean, median, mode values with concatenated dataframe using these columns "Hour','Day','Minute'.

























Hour	Minute	train_or_test	energy_mean	energy_median	energy_max	energy_min	energy_std
0	0	train	1587.161023	1569.54000	2426.0104	1078.0185	308.378350
1	0	train	1505.299102	1482.59025	2378.9524	1037.2320	304.995240
2	0	train	1462.406340	1438.71910	2362.4480	996.4836	301.973485
3	0	train	1429.134480	1394.19600	2361.4932	964.1142	307.024580
4	0	train	1428.130135	1398.27965	2389.8644	950.1786	309.482632

After processing with test data, split train and test data from concatenated data.

Model Training and Validation :

After doing all the above processing, Trained the data with various models and validated RMSE score.

Since XGB produced less RMSE and started tuning the model. PFA the results of various models tried,

Submitted at	Submission message	Score	Code File	Solution File	Final Solution
Sun, Nov-20-2022, 09:22:46 PM	XGB_2	318.44504338135	 Download	 Download	<input type="radio"/>
Sun, Nov-20-2022, 08:55:22 PM	XGB_1	306.31394450651	 Download	 Download	<input type="radio"/>
Sun, Nov-20-2022, 08:38:26 PM	XGB_updated	305.983251318779	 Download	 Download	<input type="radio"/>
Sun, Nov-20-2022, 08:16:38 PM	Modified multi model	322.488193497022	 Download	 Download	<input type="radio"/>
Sun, Nov-20-2022, 08:11:49 PM	Deep Learning	1895081.66795238	 Download	 Download	<input type="radio"/>
Sun, Nov-20-2022, 07:38:54 AM	combined_model	318.413715199885	 Download	 Download	<input type="radio"/>
Sun, Nov-20-2022, 07:14:12 AM	XGB tuned	318.244947184573	 Download	 Download	<input type="radio"/>
Sun, Nov-20-2022, 06:43:40 AM	random forest	435.27554267902	 Download	 Download	<input type="radio"/>
Sun, Nov-20-2022, 06:42:09 AM	XGB_standard scalar	446.719369347529	 Download	 Download	<input type="radio"/>
Sun, Nov-20-2022, 05:25:50 AM	xgboost	307.555382935942	 Download	 Download	<input type="radio"/>
Sun, Nov-20-2022, 05:15:41 AM	Gboost	312.829192647863	 Download	 Download	<input type="radio"/>
Sun, Nov-20-2022, 04:00:43 AM	Linear Regression Model	377.740226504796	 Download	 Download	<input type="radio"/>
