Participant Approach
0QYGSL

1. A brief on the approach used to solve the problem.
   The problem statement provided was a time-series problem. To cover the basics and test the waters, I initially started with a Machine Learning approach. I experimented with the most widely used algorithms and found XGBoost Regression and RandomForest Regression to give the highest accuracy. However, when tested against the ground truth on Analytics Vidhya, it performed poorly. Correspondingly, I moved on to the state of the art models used for time series forecasting. My experimentations revealed that the LSTM model seemed to perform exceedingly well. Therefore, I went ahead with the said model to test which of the hyperparameters gave the highest score as per the ground truth. This feedback system allowed me to me to obtain better 'future covariates' for my model and optimize other hyperperameters to create a better model with each trail. Finally, I scripted the entire process in a modular fashion to create a pipeline that could be deployed and automated for future use.

2. Which Data-preprocessing / Feature Engineering ideas really worked? How did you discover them?
   ML Approach:
   In ML approach, I initially converted the datetime column from 'object' datatype to 'datetime' datatype. Then I generated feature columns namely 'year','month','day', and 'hour' from the new datetime column. Following this, I deleted the datetime column and performed null value imputation process using KNN imputer. After that, I went ahead with test, train, and validation splitting closely followed by standard scaling to bring all the feature ranges to fit the common scale.
   DL Approach:
   The DL approach followed a similar route as per the ML approach where I initially imputed the null values using KNN imputer. Then I converted the data to a Timeseries format. Following this, I performed scaling operations on the data and generated covariates with attributes 'year','month','day', and 'hour'. After this, transformation operation was performed on the covariates. Also, the initial 'future covariates' were assumed to be identical to the past 2 years' covariates. However, with the feedback from the ground truth i.e., score on Analytics Vidhya, the energy data with higher scores was used to create later future covariates

3. What does your final model look like? How did you reach it?
   The final model is an LSTM model incorporated with 1 LSTM layer with 2% dropout, 4 attention heads, Quantile Regression, input chunk length of 48 months and forecast horizon of 24 months trained for a total of 20 epochs resulting in a MAPE of ~25% and RMSE of ~15%. Note that this is the epoch used to train the final model, however, the model was created on a trial and error basis as per the feedback from the ground truth with future covariates changing with each iteration. Hence, the cumulative epochs the model was trained for is much larger. Starting with an ML approach I finally arrived at the DL approach incorporating an LSTM model. The model is an aggregate result created from a number of trial and errors, days of hardwork, varying future covariates, ranging hyperperameters, and across a number of different free google colab accounts laying waste to hours of training time at the very last epoch. Needless to say, it was a rollercoaster of an event that left me longing for more.