Granify Data Scientist Assignment

Problem statement:
Given a dataset of energy consumed every hour for the years 2010 - 2018, arrive at an approach which will help us predict the energy requirement for a day in the future.

Your final submission should have the Python code and the Jupyter notebooks you used, your observations, approach and conclusion as a structured document.

Note: This assignment should take approximately 3+ hours

Data:
Energy consumed in MWH for years 2010 - 2018
Columns : date_and_time (yyyy-mm-dd-hh), power_MWH
Note : no additional data will be given for evaluating predictions, factor that in while doing the exercise

Data location:

https://drive.google.com/file/d/1NCl2EEvGzUfDId5drk1nY6O6ecR5XEaq/view?usp=sharing

What are we looking for?

- A structured data analysis to understand the nature of data
    - Example: is data noisy? Possible ways to overcome the problem
        - Visualization and statistical exploration of data.
        - Discuss the issues with this data that makes this problem challenging and provide evidence.
- It is **not expected** that you code each possible analytical approach. Outline the possibilities but **code only one**.
- Feature engineering - extract or create new features from the data that will help you achieve the end goal (restrict to a max of 15 features).
    - Give statistical reasoning behind the features you've selected.
- Propose and build a baseline model - need not necessarily use machine learning. Any model will be evaluated against the baseline model.

- Explain the reasoning behind selection of baseline model

• Propose a few ML models and build one against the BL - pros and cons for each explained

- Start with the simplest ML model first
- Then Increase complexity, but explain precisely what you wish to gain by it and also potentially what aspect of performance might suffer by this increase in complexity
- select one possible final machine learning model from your above ideas that you think would be best fitting for an overall solution and explain why
- Build your best likely ML model against the BL

List of Deliverables:

• Visualization and statistical exploration of data.

- Code - Python and Jupyter notebook

• Discuss the issues with this data that makes this problem challenging and provide evidence.

- Jupyter notebook with relevant statistical tests

• Feature engineering from the given data (Max 15).

- Code - Python and Jupyter notebook

• Propose architecture of a baseline model against which any parametric or machine learning model could be compared against as well as a few more complex ML models with their potential pros and cons wrt to one another and the baseline approach

- propose the design of a parametric model and walk us through the assumptions, limitations and possible ways to overcome the limitations.
- Reason for selection, limitations possible generalizations
- Discuss (not build) a few possible machine learning models that can address this problem and why
- Build and evaluate the BL model
- Choose one ML model to build against the BL model
- Evaluate the models and discuss results, conclusions and recommendations

• Your final code(s), clear reporting of the approach that you took, observations, and conclusions in a document.

- Final code should contain code for:
  - Data visualization and explorations
  - Feature engineering and
  - BL and ML models

- A report:
    - Outlining a structured approach for the whole problem.
    - A Jupyter notebook can be used for reporting consolidating all the different aspects discussed above

If you have referred to any papers, code base, etc. provide due credit in your report.