DS-203: Exercise E11

Background

- The data file e11.csv contains daily averaged values of observations logged by a data acquisition system at a chemical processing plant.
- In the data file there are 241 columns representing various parameters that are monitored and / or controlled. The data spans a period of two years and nine months.
- The parameters can be grouped into **operating** and **controllable** parameters
    - Operating parameters: these can only be monitored - they cannot be directly controlled. Examples of such parameters include compressor vibrations and the final chemical composition of the product being manufactured.
    - Controllable parameters: these are parameters that can be directly controlled. Examples include the inlet steam temperature and pressure, certain flow rates – such as cooling water, etc.
    - Following is the list of controllable parameters: c26, c27, c28, c29, c30, c31, c32, c33, c39, c139, c142, c143, c155, c156, c157, c158, c160, c161, c162, c163

The tasks

- Parameters c51, c52, c53, c54 represent vibrations of certain critical equipment. It is important to closely monitor these vibrations and keep them under control. In fact, these vibrations are so critical that:
    1. It is required to create ML models to predict the vibration levels based on the other operating and controllable parameters. These ML models will be used to raise alerts and alarms if/when they reach HIGH and CRITICAL levels respectively. The levels are defined as follows: SAFE (less than 5), MODERATE (5-10), HIGH (10-20), CRITICAL (> 20) (*all values are in appropriate units*).
    2. It is also required to create ML models to predict and control these vibrations using ***only the controllable parameters***. It is proposed to create an automated vibration control and reduction system that gets activated when the vibrations reach high and critical levels. One of the goals of this model should be to create a list of the most important parameters to change to reduce vibrations. This list should be in descending order of importance.
- c241 represents 'specific energy' (energy consumed per unit output produced) which is also a critical factor. An ML prediction model is required to understand which parameters (operating + controllable) significantly contribute to the 'specific energy', so that energy reduction research and efforts can be focused on them.
- In the context of c241, it is also required to carry out an analysis to find out the minimum number of 'independent' variables that can be used to 'only predict – not control' the specific energy consumption, and create a prediction model based on these variables.

You are required to process the available data to complete these tasks and provide your insights and recommendations based on your analysis.

Here are some questions to get you started and to guide your actions (these are, by no means, the only questions to be considered while solving the problems! You can/should create your own):

1. The usual questions first: What kind of EDA will you do on the data to get an overall understanding? How good is the data? Are there any parameters that are bad, in terms of

data not being available? What to do with such columns? Are there other columns that are not very good but which can be 'managed'? If they can be 'managed', how? Is there a need to standardize / normalize the data? Is there a need to apply any kind of data transformation to some of the parameters?

2. Will it be necessary to create ML models for each one of c51, c52, c53, c54? How to decide?
3. There are many columns in the data set – a fertile ground for conditions of multicollinearity. Should all these columns necessarily be used while training ML models? How to make this decision?
4. How to ensure the validity of the resulting ML models?

Execution and submission related information

- This exercise will be assessed for 25 marks (25% weightage) as per the following criteria:

EDA and data preparation steps, results, analysis, documentation	10 marks
ML models and solutions: creation, results, analysis, documentation	10 marks
Overall presentation style, content, quality, and effectiveness	05 marks
Predominantly LLM generated material; submissions without efforts	-10 marks

- This exercise should preferably be done in a group – of maximum three members
- Register your group using the following link:
  - https://tinyurl.com/e11-group-nov-2023
- The final submission should include:
  - PDF of a well-prepared presentation (see below)
  - PDF of the Jupyter Notebook (see below)
- Submission date: **November 12, 2023, 23:55 Hrs**.

Note:

- If you do not register your group, your submission will not be evaluated.
- **The PDF of your presentation will be the primary and the sole document for assessment**.
  - It should include all that you wish to convey to highlight to the examiners: a summary, the details, the results, the recommendations, the challenges, key achievements.
  - As frequently mentioned in class, the presentation should be crisp, should not use lengthy text, should contain adequate figures / tables – all neatly captioned, and explained where necessary.
  - Marks will be deducted if the presentation is deemed to predominantly consist of LLM created text.
- The Jupyter Notebook will constitute "evidence of effort and results".
  - It SHOULD necessarily contain ALL the images / other artifacts that you have used in the presentation.
  - It SHOULD NOT contain lengthy / debug outputs.

***