IE 48B Time Series Analytics
Final Exam

Please download the data given in the file "production_with_weather_data.zip". This file contains hourly total solar power plant production values for multiple small solar power plants around Kayseri between 2019-09-01 and 2021-12-25 (both included).

This exam is about is about providing hourly solar power prediction for the next day. This prediction is needed in the following setting. Suppose you are at day *d* and the predictions are needed for day *d+1* and you know the production values until the end of *d-1*.

Provided data file includes the weather measurements for 25 grid points (coordinates) consisting of the solar plants in the region. The relevant variables that can help for solar power plant production forecasting are described below:

TMP: Temperature at the provided location. Temperature can represent the seasonality. Moreover it is known that high temperatures affect the solar panels and decrease their efficiency.

RH: Relative humidity at the provided location. Humidity is generally high for the rainy days. Production is generally low due to the clouds when it is rainy

DSWRF: This is the short version of downward shortwave radiation flux which is known to be highly related to the production level. This is further described on the following link: https://en.wikipedia.org/wiki/Shortwave_radiation

On the other hand, I advise you not to spend too much time to understand the background information about this variable. Instead follow a data-driven approach to understand its relation to production levels.

TCDC: This is total cloud cover data (in terms of percentage) at different layers (4 layers in total). Further information about cloud types are available on https://en.wikipedia.org/wiki/List_of_cloud_types however as advised for DSWRF data, please follow a data-driven approach to understand if the variable provides information on hourly production.

You are given this weather information for 25 coordinates. This makes 175 (25x7) variables that can be helpful in the forecasting task. These are available in the file as columns in the following format: "VARIABLENAME_LATITUDE_LONGITUDE".

Your ultimate task will be to forecast hourly solar power production for the period between start of the November, 2021 until the end of the data (i.e. 2021-11-01 to 2021-12-25 both included) under the mentioned scenario (i.e. we know the production values up to *d-1* and willing to forecast the hourly levels for *d+1*) by using the regression tree-based approach (namely EBLR) covered in the class.

You can use linear regression as your base learner. Note that the method requires some fine-tuning in terms of the tree-learning parameters. I advise you to fix the depth parameter to 4 and complexity parameter to zero to save time. Please select the number of iterations for feature generation (i.e. required number of trees) by observing your performance over the iterations. If the training performance do not improve within a ratio of 1%, stop iterations.

You can evaluate your strategies based on the daily weighted mean absolute percentage error. Details of this metric is available on https://en.wikipedia.org/wiki/WMAPE. Compare your proposal with the simplest baseline of using last available data (lag 48 production → two day's ago).

Also comment on the learned decision tree model for the first 3 trees in your approach.

There are many alternative approaches to solve this prediction problem however you are expected to use the approaches covered in the class. Any other approaches (both in the preprocessing, representation) will not be accepted as a valid solution.

WHAT YOU SHOULD SUBMIT?

1- A short report that has the following sections (this part should not include any codes).

- Summary: A section that includes the summary of your approach, findings and conclusion
- Methods: A clear description of the methods evaluated. You are expected to provide the details of your data preprocessing, classification and prediction steps so that anybody who read your report can replicate your results.
- Discussion: Detailed analysis of the results and comments on your findings.

2- A script file to replicate your results described in the report. Your scripts should have comments so that a user can understand the steps taken.

Please upload your midterm as a zip file containing the report and the script file on Moodle. Name the zip file in the following format: IE48BFinal_name_surname.zip