Assignment II

General guidelines for presentations:

- You may work alone or in pair.
- Submit the report as one PDF file through Moodle course webpage or alternatively send by email to Juri Belikov (juri.belikov@taltech.ee).
- Report can be written in the free format but must contain your name(s) and student code(s) at the top of the first page.
- All the developed codes/links should be sent by email to Margarita Matson (margarita.matson@taltech.ee).

How to get data set

Each student will receive a data set (`train.csv` and `test.csv`). Files will be distributed by Margarita Matson, during Practicum 13.

Data set description

Both files include the time series of the following variables:

- `time` - timestamp
- `temp` - air temperature [in °C]
- `dwpt` - dew point [in °C]
- `rhum` - relative humidity [in %]
- `snow` - snow depth [in mm]
- `wdir` - wind direction [in degrees]
- `wspd` - average wind speed [in km/h]
- `wpgt` - peak wind gust [in km/h]
- `pres` - sea-level air pressure [in hPa]
- `price` - electricity price in Estonia on that hour [in EUR/kWh]
- `demand` - electricity demand [in kWh]

Work only with the given data sets and variables and do not look for additional sources of data. Use only methods learned during lectures and practices.

Tasks

1. Analysis [10p]:
    a. Read and understand the data.
    b. Clean the data (fill the missing data, if needed).
    c. Provide a description of the data, including statistics and at least two visuals to gain insights.
    d. Analyse distributions of each variable (except `demand`).

    - Apply any relevant transformation.

  e. Design one new feature (using **only** available features) and rank features by relevance.

    - Explain the rationale behind your ranking.

2. Modelling [15p]:

  a. The target variable is `demand`.

  b. Construct and train an autoregressive model. Use any ARMA-family model.

    - Stationarise the data, if necessary.
    - Construct, analyse, and describe ACF and PACF plots.
    - Evaluate and compare the performance using MAE metric.

  c. Construct and train a model using **at most** three additional features (exogeneous inputs, not lags).

    - Evaluate and compare the performance using MAE metric.

  d. Explain which model(s) perform better and why.

3. Forecast [15p]:

  a. Provide an out-of-sample forecast of hourly `demand` for all 7 days. Work with data set from `test.csv` file.

    - Use rolling out-of-sample forecasting approach (with 24h horizon and 0h lead time), i.e., forecast day-by-day.
    - Use any suitable forecasting strategy.
    - Use trained autoregressive model and model with additional inputs.
    - Use any simple technique (eg, naïve, drift, ACD, etc.).

These steps should help you efficiently carry out your data analysis and exploration task. For each sub-task write short explanations (a few sentences) critically describing main observations and outcomes.

If you have any specific questions or need further assistance, please feel free to ask.

Expected result

1. Textual report [40p]:

  a. PDF file with text (explanations, descriptions, conclusions, etc.), visualizations, and tables.

  b. Technical requirements: Pages: up to 10, Font size: 12pt, Font family: Times New Roman, Justification: left, Line spacing: multiple to 1,08, A4 page.

2. Source code [5p]:

  a. Provide a link to a Git repository with Python **Notebook** containing your code and description.

  b. Ensure the repository contains an IPython Notebook viewer.