Forecasting energy availability from renewable sources

January 14, 2024

1 Project Description

You're a data scientist in a company that provides energy to its clients. In the current climate of soaring energy prices, the company wants to offer a new service by which local clients that are near a source of renewable energy will have free energy when there is a surplus of energy in that area. The company wants to roll out a pilot project in which they test this scheme with a subset of customers in May 2024. As a first step, however, the company needs to have a system that can reliably predict, at least 24 hours in advance, whether there will be a surplus of energy (of either wind or solar energy) so that they can send out an alert to the customers in the area to allow them to opt-in to the slot.

In order to develop such a system, the company has provided you with access to historic records in the area (which you can download from the Assessment Information page on Moodle) from year 2000 until now and they have asked you to check the feasibility of this plan. The company is particularly worried about false positives, as this would mean that they lose money by offering free electricity to customers when there is not a surplus of energy and therefore incurring a high cost for the company.

You can find a description of what each column in the dataset represents in this website: click here.

Read the following tasks in detail and make sure you understand the project.

2 Tasks

2.1 Stage 1: Data exploration

Stage 1 is about making sure you have loaded and explored your dataset, and that you understand the data and are ready (or nearly ready) to move on to modelling using appropriate methods. Make sure you spend enough time on this assignment: cleaning and understanding the data is a big part of Data Science. Your models won't work if they are not appropriate for the problem at hand.

Check the marking scheme for the 'Data Exploration' assessment. It's available on Moodle in the Assessment Information page – more details are in Section 3 of this document. Based on this marking scheme, the main tasks you should complete are:

1. Load and explore the data set, leaving a subset of data separate from the exploration to avoid overfitting.
2. Clean and preprocess the data set. By the end of Stage 1, you should have a clean and cohesive data set.
3. Do research on adequate thresholds that would indicate surplus of energy — this will have to be from external sources, not from your dataset only. Make sure you keep a record of values and the references you used to find these thresholds, as the marking scheme explicitly asks for them.

Note that you are NOT asked to perform any modelling at this stage. During your lab demo, we will check that you understand the project and have a plan for how to analyse and model the data for the second stage that is reasonable and feasible.

2.2 Stage 2: Modelling and testing

In Stage 2 your task is to complete the project described in Section 1 of this document. In Stage 1 you understood and prepared your data for modelling. In Stage 2 you need to answer the question posed to you by the energy company and present your results to the CEO.

Check the marking scheme for the Stage 2 assessments ('Final project Code' and 'Final project Demo'). They can be found in the Assessment Information tab on Moodle and in Section 3 of this document. Based on this marking scheme, the main tasks you should complete are:

1. Some form of modelling to answer the project that the company tasked you with. The modelling is up to you, as long as it follows good practice in data science (i.e., using cross-validation and train/test splits properly and as appropriate for this data set, problem, and the chosen modelling approach). As this is a research-led module, you are encouraged to check the available literature and find out what has worked in the past, but you need to present something new/different as part of your project.

2. While you're modelling, reflect on the project description given above (Section 1 of this document) and make sure you pay special attention to the requirements set out by your manager.

3. For your presentation, make sure that all your decisions are justified and that you present your findings clearly and concisely. Any assumptions made and references used must be stated. As a data scientist, you must reflect on your results. Talk about your findings and what they indicate. Finally, you **must** answer the question: are you confident that your model fulfils the requirements for the company? Is it ready to be deployed in May 2024? Are there any limitations they should consider? Do you have any other insights from the data that can help the company in the future (e.g., for future features that can be commercialised)?

3 Deliverables and marking criteria

3.1 Stage 1: Data exploration

All the code used must be submitted to FASER in a **zip** folder.

Your FASER submission must include:

- A README file with a description of the project and instructions on how to use/run the code. Any assumptions made must also be stated in the README file.

- The code that you used to carry out the exploration. One Jupyter notebook is enough, but it should not be just a stream of figures/plots with no justification of why they're informative: make sure your notebook is properly documented and there are useful comments in the code. Explain the insights you get from each figure. There should be no errors/warnings in the notebook/s, and try to avoid repeating the same code multiple times (this is what functions are for). Headings for different subsections are a bonus.

3.1.1 Marking criteria for Data exploration [20% of final mark]

The following aspects will be assessed about the code and in the labs:

- Is there evidence of data loading? [5%]
  - Has the data been loaded?
  - How much data was loaded? (in terms of files and locations)
- Is there sufficient appropriate exploration? [25%]
  - Have the inputs/features been properly explored according to the type of data?
  - Do plots have labels and legends?
- Data preprocessing: Has the data been cleaned properly? [25%]
  - Has the data been preprocessed properly, using adequate methods for the type of data that is being considered (including data splits)?
  - Is there a coherent final dataset?
- Assumptions and Research: Have the assumptions been stated? [25%]
  - Has the student used any references?
  - Has the student stated any assumptions made in order to carry out the project?
  - What type of assumptions are there? (Only for some types of energy, or more generally also about the catchment area?)
- Notebook [20%]
  - Is the notebook well presented?
  - Are there meaningful comments?

  - Is the code free of data dump and errors?
  - Does the notebook have structure?
  - Are plots explained?

Note that you will only get your mark for the Data Exploration assignment if you present your work during the lab session. Failure to attend will result in an automatic mark of 0. Failure to explain the code and discuss your results will also result in an automatic mark of 0.

3.2 Stage 2: Modelling and testing

There are two deliverables for Stage 2, all of which must be submitted to FASER:

3.2.1 Final project Code

All the code used must be submitted to FASER in a **zip** folder.
Your FASER submission must include:

- A README file with a description of the project and instructions on how to use/run the code. Any assumptions made must also be stated in the README file.
- All the code that is necessary to go from the original dataset that was given to you until your final results.
  - The notebook with exploration and preprocessing (this can be two notebooks, but there shouldn't be lots of repetition across them).
  - The notebook used for modelling and to obtain the final results.

  Exploration/preprocessing and modelling MUST be in different jupyter notebooks. Make sure your notebooks are properly documented and there are useful comments in the code. Explain the insights you get from each figure, and justify your methods through markdown cells and comments. There should be no errors/warnings in the notebooks, and try to avoid repeating the same code multiple times (this is what functions are for). Headings for different subsections are a bonus.

3.2.2 Final project Demo

The final project demonstration deliverable is submitted through FASER in the form of a presentation. This presentation will be seen by the CEO of the company, so you should focus on a high level description of what you did and mostly on your insights, results, and what it means for the company. Make sure you answer the questions in point 3 of Section 2.2 — what's the bottom line for the company?
Your FASER Submission must include:

- The document used to present your results.
- A 10-minute video presentation using the slides submitted. If the video is longer than 10 minutes, only the first 10 minutes will be watched and evaluated.

You must submit both the video and the presentation or you will receive a mark of 0.

3.2.3 Marking criteria for Final project Code [31% of final mark]

The following aspects will be assessed about the code and in the labs:

- Is the preprocessing/cleaning correct? [15%]
  - This is a chance for you to fix any errors that were identified in the Data Exploration assignment
  - Has the data been preprocessed properly, using adequate methods for the type of data that is being considered (including splitting the dataset)?
  - Have errors from Stage 1 been fixed (if any)?
  - Is there a coherent final dataset appropriate for the Final project assignments?
- Assumptions and Research: Have the assumptions been stated? [10%]
  - Has the student stated any assumptions made in order to carry out the project?

  - What type of assumptions are there? (e.g., only for some types of energy, or more generally also about the catchment area?)
  - Has the student used any references?
  - Are the assumptions reasonable?
- Are Machine learning and Data Science conventions followed in the pipeline? [25%]
  - Are the methods appropriate for the project? — including data splits, type of modelling, metrics used, etc.
  - Are there errors in the pipeline/s?
- How well organised is the code? [10%]
  - Are the notebooks well presented?
  - Are there meaningful comments?
  - Is the code free of data dump and errors?
  - Does the notebook have structure?
  - Are plots explained?
- Is there a README file describing the project and organisation of the code? [5%]
- How much depth was achieved in the project? [20%]
  - How many types of energy were used (solar only/wind only/both)?
  - How many locations were used (Colchester/Brighton/both)?
  - Has a decision-making system been implemented?
  - Is there a reflection/analysis of false positives?
  - Do the final conclusions reflect the results?
- Has the student discussed the project with the Module Supervisor? [15%]
  - This discussion can take place over office hours or in labs, but must be face to face and before the deadline for this assignment.

3.2.4 Marking criteria for Final project Demo [31% of final mark]

The following aspects will be assessed about the code and in the labs:

- Introduction: Is the project introduced correctly? [10%]
  - Please include 1–2 slides introducing the project and main objectives
- Methods: Are they summarised? [10%]
  - The methods must be stated at a high level — remember that you're not presenting to data scientists, but to the CEO of the company. Just a short summary of methods is enough
- Results: Are they presented coherently? [15%]
  - Has the student made appropriate use of graphs and plots to summarise the results?
- Assumptions and Research: Have the assumptions been explained? [15%]
  - Has the student used any references (please place them at the end of the presentation, in a separate slide)?
  - Has the student stated any assumptions made in order to carry out the project?
  - What type of assumptions are there? (Only for some types of energy, or more generally also about the catchment area?)
  - Are the assumptions appropriate?
- Conclusions and recommendations [20%]
  - Are there conclusions for the CEO?

  - Do the conclusions match the results presented?
  - Is there a final recommendation for the company?
  - Does the recommendation match the results presented?

- Presentation style [10%]
  - Is there a title page with a catchy title and the registration number of the student?
  - Is the presentation coherent?
  - Are the slides pleasing to the eye?
  - Is there a good balance of text and figures/plots? Are plots of good quality?

- Are the project questions sufficiently answered? [20%]
  - How many types of energy were used (solar only/wind only/both)?
  - How many locations were used in the analyses?
  - Has a decision-making system been implemented?
  - Is there a reflection/analysis of false positives?

- Coherence between Demo and Code [up to -20%]
  - This section checks for consistency between the results presented in the presentation and those submitted in the code separately. If results don't match, you might lose up to 20% of the marks for this assignment.

    The figures and numbers from your presentation must match those from the code.

    But we will also be looking at other general inconsistencies.