Assignment 2: Data Science Project

March 14, 2024

Due Date

Sunday 14th April, 11:59PM

Late submissions

No extensions are allowed for this assignment due to the oral component of this assignment during Week 12.

Hypothetical Scenario and Objective

As a team of data scientists, you have been tasked to develop a model that can be used to aid energy companies in understanding and planning for energy consumption. First your team needs to ensure the data is ready for analysis. After that, your goal is to develop a model that predicts the maximum daily price category based on the provided data.

The following two datasets are provided:

- `weather.csv` contains key weather indicators, such as minimum and maximum temperatres for the city of Melbourne for each day between March 2023 and 9 August 2023. This data has been extracted from the Bureau of Meteorology and collated into a single file for your convenience.

- `price_and_demand.csv` contains energy price and demand figures for the state of Victoria at a half an hour interval between March 2023 and 9 August 2023. This data has been extracted from the Australian Energy Market Operator (AEMO), and collated into a single file for your convenience. The file has been modified for the purposes of this assignment. You may also opt to download and inspect the raw data from AEMO for the period of March 2023 and 9 August 2023 and consider this for this assignment if you wish.

Working in a team of 3-4 students, you should:

- Perform data cleaning on the provided data.

- Develop a model which predicts the maximum daily price category based on the provided data.

- Evaluate the effectiveness of your model.

- Present analysis on how weather data affects price and demand.

- Discuss the limitations of your work and ways these could be potentially addressed.

- Write a report about your exploratory analysis supported with tables and figures.

- Present a summary of your project during Week 12 live webinar and demonstrate your individual understanding of your data analysis.

Experimental Design and Team Responsibilities

With the files provided to you, you might find that building a classification model is more appropriate. But if you would also like to build a regression model, you may download the raw data from AEMO. Your report should present the process of building your model i.e. selecting features included in your model and fine-tuning the hyperparameters. It is your responsibility to assign roles to each team member and ensure that all team members contribute equally to your project.

On Week 12, one or two members will present a 2-3 minute summary of your project. Then each individual member of the group will be asked 1-2 questions to assess their understanding of the project. Each member is expected to know the project well. The questions will be specific to your team's code and report. Each member will be marked individually on the quality of responses to the questions.

Assessment

Your Report

Your report should be written in a Word document (.pdf or .docx). It must display a word count and should be 2000-2500 words in length. The word count limit excludes any figures, tables, and appendices. Anything written over the word limit will not be assessed. Your report should include the following information:

1. What data cleaning methods have you applied? Why have you chosen these methods over other alternatives? Give examples to support your chosen methods.

2. Explain the process of building your best model. How did you select the features included in the model? How does your model work?

3. How effective is your model? How have you evaluated this?

4. What insights about weather and daily energy usage can you draw from your analysis? Discuss any significant results.

5. What are the limitations of your results and potential ways to address these limitations?

 Your report should also include the following:

1. Make effective use of visualisations to support your arguments.

2. A section for references to list any resources you have used to exhibit your academic honesty.

Git repository (Optional)

All of the code you develop as part of this project is encouraged to be stored in a GitHub repository. Ensure that this is set as a private repository since this is an assessment.

Only one member of your group should create a GitHub repository, the other group members should be added to the same GitHub repository. This will ensure that all group members are able to collaborate on the same codebase.

README file for Git repository

A README file must be included in your git repository with sufficient information on how to run the code to get the results, including the random_state values if required. You must ensure that this README file contains the full names of each member of your group. Important note: Assessors must be able to replicate the models based on your explanations, provided code, and instructions provided in this README file.

Submission Instructions

1. Your final report must be uploaded via Canvas by the due date under "Group Assignment - Report Submission".

2. All of your code files, and any other supporting files used, should be placed in a .zip archive and uploaded via Canvas by the due date under "Group Assignment - Code Submission.".

3. Any numerical results or visualisations used in the final report must be reproduced by running your code.

4. You must also include a link to your GitHub repository if you have used it.

Academic Integrity

You are expected to follow the academic integrity guidelines on the University website
https://academicintegrity.unimelb.edu.au/

Table 1: Marking Rubric

Requirement	Description	Maximum Mark
Data cleaning and visualisation	What data cleaning methods have you applied? Why have you chosen these methods over other alternatives? Give examples to support your chosen methods.	9
Building a Model	Explain the process of building your model. How does your model work?	6
Model effectiveness and evaluation	How effective is your model? How have you evaluated this?	6
Data Analysis	What insights about weather and daily energy usage can you draw from your analysis? Discuss any significant results.	6
Limitations	What are the limitations of your results?	5
Individual Quality of Reponses to Questions	Each group member will be asked 1-2 questions during the Week 12 live webinar. You will be marked based on the quality of your responses to the questions.	8
Overall Maximum Mark		**40**