BUSS6002 Assignment 2

Semester 1, 2023

Due Date

- Due: at 12:00pm (noon) on Monday, 22nd May 2023 (start of week 13).
- A late penalty of 5% per day applies if you submit your assignment late without a successful special consideration or simple extension. See the Unit Outline for more details.
- You may submit multiple times but only your last submission will be marked.

Instructions

- You must submit:
  - a **written report** (PDF) with the following filename format, replacing `STUDENTID` with your own student ID: `BUSS6002_A2_STUDENTID.pdf`.
  - a **Jupyter Notebook** (`.ipynb`) file with the following filename format, replacing `STUDENTID` with your own student ID: `BUSS6002_A2_STUDENTID.ipynb`.
- There is a limit of 2000 words for your report (excluding equations, tables, and captions).
- All plots, computational tasks, and results must be completed using Python.
- Each section of your report must be clearly labelled with a heading.
- Do not include any Python code as part of your report.
- All figures must be appropriately sized and have readable axis labels and legends (where applicable).
- The submitted `.ipynb` file must contain all the code used in the development of your report.
- The submitted `.ipynb` file must be free of any errors, and the results must be reproducible.

Rubric

This assignment is worth 20% of the unit's marks. The assessment is designed to test your technical ability and statistical knowledge in modelling a real-world dataset, as well as your communication skills in writing a concise and coherent report presenting your approach and results.

Assessment Item	Goal	Marks
Section 1	Introduction	3
Section 2	Candidate models	10
Section 3	Model estimation and selection	12
Section 4	Model evaluation	8
Section 5	Conclusion	3
Overall Presentation	Clear, concise, coherent, and professional	4
Total		40

Table 1: Assessment Items and Mark Allocation

Overview

The Australian federal government is building a website to provide households with a tool to determine if installing a rooftop solar panel system is right for them. On the website, users will be able to enter information about their house and the website will provide an estimate of the possible solar power generation.

The government collected a random sample of existing households with solar panels, including information about the households, solar panel installation and the associated solar power generation. The generation from the solar panels was collected from 1/1/2022 to 31/12/2022.

The data has been assembled from a multiple sources including the customer energy retailer, energy distributors and solar installers. Sampling is limited to installations with:

- a single solar panel array or multiple arrays that are oriented identically,
- rooftop installation only.

After you presented your EDA from Assignment 1, you have been given a new task: determine if a model can be built to predict `Generation`, that can outperform simple baselines.

Data Files

The following files are available on Canvas.

File	Description
SolarSurvey.csv	Data file with 3000 observations.
DataDictionary.txt	Data dictionary containing description of each variable

Table 2: Files Provided

1 Introduction

In this section of your report, you should

- provide a brief project background so that the reader of your report can understand the general problem that you are solving;
- state the aim of your project;
- briefly describe the dataset;
- briefly summarise your key results.

2 Candidate model

Propose at least three candidate models for predicting the response variable `Generation`. For $i \in 1, 2, 3$, each candidate model should take the form

$$y = f_i(\mathbf{x}_i, \boldsymbol{\beta}_i) + \epsilon_i$$

where y is the `Generation`, and $\mathbf{x}_i$, $\boldsymbol{\beta}_i$, and ϵ_i are the predictor vector, parameter vector, and the error term of the i-th model, respectively. The set of variables chosen for the feature vector $\mathbf{x}_i$ should be a subset (or constructed from a subset) of the predictors in the provided dataset. You may label your models M1, M2, and M3. The proposed models should be different in terms of model complexity (i.e., number of parameters) and/or feature engineering.

For each proposed model, you should:

- clearly define the function f_i, which can be either linear or nonlinear with respect to x_i;
- clearly define the feature vector $\mathbf{x}_i$;
- justify your choices of fi and $\mathbf{x}_i$;
- state any assumptions on the error term ϵ_i;
- discuss how the model parameters β_i can be estimated.

Hints:

- one effective way to motivate/justify your choices of f_i and $\mathbf{x}_i$ is to present the relevant evidence in the data.
- carefully consider how the predictors are related to the target

3 Model estimation and selection

Select the best model from the set of candidate models proposed in Section 2 using the "validation set" approach. In this section of your report, you should:

- include a description of the model selection procedure that you adopted;
- report and discuss the estimation results (based on the training set) of each candidate model;
- discuss whether each candidate model is correctly specified based on residuals (obtained from fitting each model to the training set);

- report the validation performance (MSE) of each candidate model;
- identify the best model;
- discuss the complexity of the selected model in terms of bias-variance tradeoff.

The description of the model selection procedure (first point above) should provide enough details so that the reader is able to implement exactly what you have done by following your description.

4 Model evaluation

Evaluate the generalisation performance of the selected model in Section 3 against two benchmark models. The generalisation performance should be measured by the observed MSE calculated using the test set.

In this section of your report, you should

- combine the training and validation sets and re-estimate the selected model on the combined set;
- describe the model evaluation procedure;
- describe the two benchmark models;
- report and discuss the generalisation (i.e., test set) performance of the selected model against the two benchmark models.

The two benchmark models are specified in the following subsections.

4.1 Benchmark Model 1

The **Benchmark Model 1** (BM1) predicts the `Generation` by averaging the observed `Generation` values for modern systems within each city. Modern systems are those installed in 2019, 2020 and 2021.

Let D be the set constructed by combining (or concatenating) the observed `Generation` in the training and validation sets. Let $C(x)$ be the subset of D that contains only the `Generation` from the city of x installed between 2019 and 2021. For example C('Sydney') contains the `Generation` in D from Sydney only. Then BM1 is given by:

$$\hat{y} = \frac{1}{m(x)} \sum_{y \in C(x)} y$$

where $m(x)$ is the size of the set $C(x)$.

4.2 Benchmark Model 2

The **Benchmark Model 2** (BM2) extends BM1 by further grouping based on panel capacity i.e. let $C(x_1, x_2)$ be the subset of D that contains only the `Generation`:

- from the city of x_1,
- with the panel capacity of x_2
- installed between 2019 and 2021.

5 Conclusion

In this section of your report, you should

- discuss your findings;
- discuss any limitations of your project;
- suggest any potential extensions for future work.