BUSS6002 Assignment 2

October 10, 2022

Instructions

- Due: at 23:59 on Friday, October 28, 2022 (end of week 12).
- You must submit a **written report** (in PDF) with the following filename format, replacing `STUDENTID` with your own student ID: `BUSS6002_A2_STUDENTID.pdf`.
- You must also submit 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 appr
  applicable).
- The submitted . f your report.
- The submitted `.ipynb` file must be free of any errors, and ible.
- You may submit multiple times but only your last submissi
- A late penalty applies if you submit your assignment late without a successful special consideration. See the Unit Outline for more details.

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

Being able to accurately predict the sale prices of residential properties is crucial to many aspects of the economy. Some compa ts with predictions of property sal **uild a model to predict sale prices** using data on Y ited States. The dataset contains sale prices between 2006 and 2010 of all resid Ames, as well as many numerical and categorical features (i.e., variab dwelling. The following downloadable files are available on Canv

File	Description
`AmesHousing.txt`	Data file containing 2,930 observations and 82 variables
`DataDocumentation.txt`	Data dictionary containing description of each variable
`AmesResidential.pdf`	A map of Ames

Table 2: Files Provided

Data

Place the data file `AmesHousing.txt` in the same location (i.e., directory) as your Jupyter Notebook file (`.ipynb`), and then read the data into a `pandas DataFrame` object using *exactly* the following code.

```
import pandas as pd

data = pd.read_csv(
    'AmesHousing.txt',
    sep='\t',
    keep_default_na=False,
    na_values=[''])
```

1 Introduction

In this section, 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 models

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

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

where y is the sale price of a property, 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 81 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). For each proposed model, you should:

- clearly define the func ect to $\mathbf{x}_i$;

- clearly define the feat
- justify your choices of f_i and $\mathbf{x}_i$;

- state any assumptions on the error term ε_i;
- discuss how the model parameters $\boldsymbol{\beta}_i$ can be estimated.

Hint: one effective way to motivate/justify your choices of f_i and $\mathbf{x}_i$ is to present the relevant evidence in the data.

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, 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. The two benchmark models are specified as follows.

- Let C be the set constructed by combining (or concatenating) the observed sale prices in the training and validation sets. The **first benchmark model** (BM1) is the "constant mean" model given by

$$\hat{y}_{\text{BM1}} := \frac{1}{m} \sum_{y \in C} y,$$

  where $m > 0$ is the size of the set C. That is, BM1 will always give the sample mean of C as its prediction, regardless the values of any predictors.

- Let $N(x)$ be the subset of C that contains only the sale prices from the neighbourhood x. E.g., $N(\texttt{'OldTown'})$ contains the sale prices in C that are associated with the neighbourhood 'OldTown'. The **second benchmark model** (BM2) is the "neighbourhood mean" model given by

$$\hat{y}_{\text{BM2}} := \frac{1}{m(x)} \sum_{y \in N(x)} y,$$

  where $m(x) \leq m$ is the size of the set $N(x)$. That is, BM2 predicts the sale price by the average price of the corresponding neighbourhood.

In this section, you shoul

- combine the trainin e 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.

5 Conclusion

In this section, you should

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