COLEG GWYDDONIAETH A PHEIRIANNEG
COLLEGE OF SCIENCE AND ENGINEERING

YSGOL CYFRIFIADUREG A PHEIRIANNEG
SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

PRIFYSGOL BANGOR UNIVERSITY

ASSESSMENT BRIEF
BRIFF ASESU

Module Title, Code, Year	ICE-4006-Data Science, 2023-24
Assessment Name, Length, Duration	Half Term, 2000 words (+ or- 10%) which does not include referencing, graphs, and tables.
% Contribution to Module Mark	This assessment contributes to 40% of the overall module mark
Deadline (Date and Time) for Submission	13 Nov 2023, 23:59
Format/Location of Submission	Assignments are to be submitted through Turnitin in Blackboard.
Prepared by:	Dr Vahid Seydi

Module Learning Outcomes

- Employ data science techniques with a data-set.
- Report results of experiments analysing data.
- Evaluate the effacacy of experiments conducted.
- Examine appropriate methods to interpret and visualize data.

Plagiarism and Unfair Practice

Plagiarised work will be given a mark of zero. Remember when you submit you agree to the standard agreement: This piece of work is a result of my own work except where it is a group assignment for which approved collaboration has been granted. Material from the work of others (from a book, a journal or the Web) used in this assignment has been acknowledged and quotations and paraphrasing suitably indicated. I appreciate that to imply that such work is mine, could lead to a nil mark, failing the module or being excluded from the University. I also testify that no substantial part of this work has been previously submitted for assessment.

Late Submission & Extensions

Work submitted within one week of the stated deadline will be marked but the mark will be capped at 40%. **A mark of 0% will be awarded for any work submitted 1 week after the deadline.**
Acceptable reasons for submitting work late include: Serious personal illness with a doctor's certificate (a self-certified medical note should not be accepted); The death of a relative or close friend; Serious family problems such as divorce, separation, and eviction. Examples of unacceptable reasons for failing to submit work on time include: Having exams; Having other work to do; Not having access to a computer; Having computer related problems; Being on holiday; Not being able to find information about a subject. **Contact the Lecturer if you need an extension.**

Marking Scheme

Please remember that marks are provisional until they are confirmed by a board of examiners. The marking scheme will be detailed in the requirements document for each exercise.

excellent (≥70%): Assemble critically evaluated, relevent areas of knowledge and theory to constuct professional-level solutions to tasks and questions presented. Is able to cross-link themes and aspects to draw considered conclusions. Presents outputs in a cohesive, accurate, and efficient manner.

good (≥60%): Is able to analyse a task or problem to decide which aspects of theory and knowledge to apply. Solutions are of a workable quality, demonstrating understanding of underlying principles. Major themes can be linked appropriately but may not be able to extend this to individual aspects. Outputs are readily understood, with an appropriate structure but may lack sophistication.

threshold performance (≥50%): Uses key areas of theory or knowledge to meet the Learning Outcomes of the module. Is able to formulate an appropriate solution to accurately solve tasks and questions. Can identify individual aspects, but lacks an awareness of links between them and the wider contexts. Outputs can be understood, but lack structure and/or coherence.

below threshold performance (< 50%): An attempt has been made yet does not address considerable areas of the criteria.

Assessment Feedback

Formative (On-going): Verbal Feedback – Verbal feedback will be available by request at each lecture/workshop. It is suggested that you keep a written note of this feedback to aid in your personal development. You will also have a short meeting to discuss your design after you have submitted it. (Instant)

Summative (Post Assessment): Written Feedback – Written feedback will be made available through blackboard after an assignment is submitted. To access your written feedback see the comments section of your assignment submission.(1-2 weeks)

Referencing

The school uses the IEEE referencing style: IEEE-Reference-Guide.pdf

Detailed Assessment Guidance

Aim:

This half-term exam is designed to assess your understanding of key concepts covered in the first five weeks of the data science module. You will apply data science techniques and methodologies to solve a real-world problem related to house price prediction. The exam will test your ability to work with data, apply regression analysis, optimize models, assess model performance, handle bias and variance, and effectively communicate your findings.

Instructions:

1. Dataset Description:
You will be provided with two house price prediction datasets. The first dataset is identical to the one used in practical sessions (labeled dataset), which you will use for sections A to E. The second dataset has the target variable, the house price, removed (unlabeled dataset), and this one will be used for section F.

2. Methodology:
Utilize the knowledge and skills you have gained during the first five weeks of the module, including data preparation, simple regression, multiple regression, feature selection, regularization (ridge and lasso), model assessment (train/true/test errors), data splits, optimization, and managing bias and variance.

3. Deliverable:
Prepare a comprehensive Jupyter Notebook that encompasses the following sections:

A. Data Exploration (Approx. 100 words - 5 marks)

A1. Load the labelled dataset.
A2. Provide a summary of the dataset, including descriptive statistics and data visualizations to gain insights into the data.

B. Data Pre-processing (Approx. 100 words - 5 marks)

B1. Remove categorical features and perform any necessary pre-processing steps.
B2. Utilize the `sklearn.model_selection.train_test_split` method to split the data into training and test sets. Set the `random_state` to the last two digits of your student ID. The specific `train_size` is arbitrary. These two sets will serve as the foundational datasets for all subsequent sections.

C. Model Building (Approx. 800 words - 40 marks)

For each of the following sections, employ k-fold cross-validation to further split your training data into training and validation subsets as needed.

C1. Model 1: Simple Regression
- Implement simple regression model using one selected input.
- Optimize the model using both the closed-form and gradient descent approaches.

C2. Model 2: Polynomial Regression
- Visualize the relationship between price and the chosen input.
- Determine and justify the set of features for polynomial regression, such as 1, x, x^2, or $log(x)$.

C3. Model 3: Multiple Regression
- Apply forward selection for feature selection.
- Implement multiple regression with a reasonable number of selected features.

C4. Model 4: Ridge Regression
- Create a set of alpha values for tuning your model and select the best one.

C5. Model 5: Your Suggested Model
- Present your suggested model.

D. Explanation Following Concepts (Approx. 600 words - 30 marks)

D1. Explain the learning rate selection process for Model 1 when optimizing model parameters using gradient descent. Describe the impact of using a large learning rate and the advantages of starting with a large rate and reducing it iteratively.
D2. Analyse the expected Residual Sum of Squares (RSS) in Model 1 with respect to the optimization method used (GD or closed-form) and discuss whether it represents the minimum point for RSS?
D3. Discuss the indicators that help determine if Model 2 is not overfitted.
D4. Explain how you select the best alpha for Model 4.
D5. Explain the rationale behind your suggestion for Model 5.
D6. Evaluate the complexity of the five models and explain which one has the highest bias/variance and why, and which one has the least bias/variance and why.

E. Model Assessment (Approx. 200 words - 10 marks)

E1. Predict the test data using all five models and calculate the mean squared error (MSE).
E2. Determine which model performs the best and provide an explanation.

F. Model Suggestion for Unlabelled Data (Approx. 200 words - 10 marks)

F1. Recommend one of the five models for application to unlabelled data and elucidate the reasons behind your choice.
F2. Apply your selected model to the unlabelled dataset, create a "predicted_price" column, save the results to a CSV file, and submit both the CSV file and your Jupyter Notebook. Please ensure that the CSV file is named after your student ID.

Note: The word count approximations are provided for guidance. Marks will be awarded based on the quality and depth of your explanations and analysis.