Department of Computing and Mathematics

ASSESSMENT COVER SHEET 2024/25

Module Code and Title:	6G7V0017 Advanced Machine Learning
Assessment Set By:	L Gerber, I Chatterjee
Assessment ID:	1CWK100
Assessment Weighting:	100%
Assessment Title:	Machine Learning Project
Type:	Individual
Hand-In Deadline:	See Moodle
Hand-In Format and Mechanism:	Moodle (a single .zip file with pdf report and ipynb notebook)

Learning outcomes being assessed:

1. Explain theoretical and practical aspects of popular, state-of-the-art supervised and unsupervised machine learning techniques and models.

2. Select, implement, apply, optimise, and evaluate suitable machine learning algorithms for given domains for producing a range of models from data.

3. Critically analyse the merits and limitations of machine learning approaches, techniques, and models with respect to their technical properties as well as their impact on stakeholders and society at large.

4. Apply a wide range of transferable skills and attributes applicable to industry and research.

Note: it is your responsibility to make sure that your work is complete and available for marking by the deadline. Make sure that you have followed the submission instructions carefully, and your work is submitted in the correct format, using the correct hand-in mechanism (e.g., Moodle upload). If submitting via Moodle, you are advised to check your work after upload, to make sure it has uploaded properly. You should make at least one full backup copy of your work.

Penalties for late submission

The timeliness of submissions is strictly monitored and enforced.

All coursework has a late submission window of 7 calendar days, but any work submitted within the late window will be capped at 50%, unless you have an agreed extension. Work submitted after the 7-day late window will be capped at zero unless you have an agreed extension. See ‘Assessment Mitigation’ below for further information on extensions.

Please note that individual tutors are unable to grant any extensions to assessments.

Assessment Mitigation

If there is a valid reason why you are unable to submit your assessment by the deadline you may apply for Assessment Mitigation. There are two types of mitigation you can apply for via the module area on Moodle (in the ‘Assessments’ block on the right-hand side of the page):

- **Non-evidenced extension**: does **not** require you to submit evidence. It allows you to add a **short** extension to a deadline. This is not available for event-based assessments such as in-class tests, presentations, interviews, etc. You can apply for this extension during the assessment weeks, and the request must be made **before** the submission deadline. For this assessment, the non-evidenced extension is 2 days.
- **Evidenced extension**: requires you to provide independent evidence of a situation which has impacted you. Allows you to apply for a longer extension and is available for event-based assessment such as in-class test, presentations, interviews, etc. For event-based assessments, the normal outcome is that the assessment will be deferred to the summer reassessment period.

Further information about Assessment Mitigation is available on the dedicated Assessments page.

Plagiarism

Plagiarism is the unacknowledged representation of another person's work, or use of their ideas, as one's own. Manchester Metropolitan University takes care to detect plagiarism, employs plagiarism detection software, and imposes severe penalties, as outlined in the Student Code of Conduct and Academic Misconduct Policy. Poor referencing or submitting the wrong assignment may still be treated as plagiarism. If in doubt, seek advice from your tutor.

As part of a plagiarism check, you may be asked to attend a meeting with the Module Leader, or another member of the module delivery team, where you will be asked to explain your work (e.g. explain the code in a programming assignment). If you are called to one of these meetings, it is very important that you attend.

Use of generative AI

The use of generative AI is permitted in this assessment, so long as it is used in accordance with the instructions provided in the 'Are you allowed to use AI in assessments?' section of the AI Literacy Rise Study Pack. All submitted work must be your own original content.

If you are unable to upload your work to Moodle

If you have problems submitting your work through Moodle, you can send your work to the Assessment Management Team using the Contingency Submission Form. Assessment Management will then forward your work to the appropriate person for marking. If you use this submission method, your work must be sent **before the published deadline**, or it will be logged as a late submission.

Assessment Regulations

For further information see the Postgraduate Assessment Regulations on the Assessments and Results information pages.

Formative Feedback:	Formative feedback on provisional work will be given primarily verbally in lab sessions and also on 1-to-1 sessions during office hours arranged with the unit lecturers Luciano Gerber (L.Gerber@mmu.ac.uk) and Indranath Chatterjee (I.Chatterjee@mmu.ac.uk).
Summative Feedback:	Summative feedback is provided with a breakdown of marks for the assessment criteria and a general feedback document via Moodle.

Coursework Specification (1CWK100)

6G7V0017 Advanced Machine Learning

Luciano Gerber, Indranath Chatterjee

24/25 Semester 2

Overview

In this assignment, students consolidate their learning in the Advanced Machine Learning unit by addressing specific machine learning tasks through the creation of a working solution for a real-world regression problem. You will use the **Car Sale Adverts dataset** provided by **AutoTrader (AT)**. The dataset includes anonymised vehicle information used to predict selling prices. The tasks emphasise automated feature selection, tree ensembles, ensemble methods, model interpretability (SHAP/PDP), dimensionality reduction, polynomial regression, and clustering techniques for feature engineering.

You will submit a structured PDF report, accompanied by a documented and reproducible Python notebook (.ipynb), as a single `.zip` file via Moodle.

Obtaining and Exploring the Dataset

The dataset provided by AutoTrader is available in our Moodle area here. It contains approximately 400K rows.

To manage computational resources effectively, consider:

- Using a sample of the dataset;
- Visualising large datasets with heatmaps/hexbins instead of scatter plots;
- Reducing high-cardinality categorical features before encoding.

Tasks to Perform

Your tasks are divided into two main parts as follows:

Preliminary Task (10%)

1. **Data Description and Pre-Processing** (e.g., identify and handle missing values, outliers, noise; encode categorical variables; scale features; partition data into training, validation,

and test sets). You will need to describe the **output dataset** of this task which is assumed otherwise to be the input of remaining tasks.

Part I (50%)

2. **Automated Feature Selection** (10%): Apply and evaluate methods such as recursive feature elimination, making choices and justifications.
3. **Tree Ensembles** (10%): Train, tune, and evaluate tree ensemble models like Random Forests and Boosted Trees.
4. **Ensemble of Tree Ensembles** (10%): Combine multiple tree-based models to enhance prediction accuracy (e.g., stacking, averaging).
5. **Feature Importance** (10%): Obtain, analyse, and interpret feature importance metrics via permutation importance and SHAP.
6. **SHAP/PDP Model Explanations** (10%): Use SHAP values and Partial Dependence Plots to interpret and explain your models.

Part II (40%)

7. **Dimensionality Reduction (Linear)** (10%): Apply and analyse methods like PCA.
8. **Dimensionality Reduction (Non-Linear)** (10%): Implement and evaluate methods such as Isomap.
9. **Polynomial Regression** (10%): Build and assess polynomial regression models.
10. **Clustering for Feature Engineering** (10%): Perform clustering techniques (e.g., K-Means) for deriving new features.

Marking Criteria

The **assessment criteria** is based on the University's PGT Assessment Criteria and stepped marking, and includes aspects of code/explanations such as:

- clarity, conciseness, style, correctness
- usefulness, challenge, creativity, initiative
- efficiency, reusability, and generality

Report Components and Weights

Each component should include relevant code snippets, outputs (tables or plots), and concise explanations (one or two paragraphs). Limit each component to approximately one page (±20%):

Component	Weight	Limit
1. **Data Description and Pre-Processing**	10%	1 page
Part I	50%	5 pages
2. Automated Feature Selection	10%	1 page
3. Tree Ensembles	10%	1 page
4. Ensemble of Tree Ensembles	10%	1 page
5. Feature Importance	10%	1 page
6. SHAP/PDP Model Explanations	10%	1 page
Part II	40%	4 pages
7. Dimensionality Reduction, Linear	10%	1 page
8. Dimensionality Reduction, Non-Linear	10%	1 page
9. Polynomial Regression	10%	1 page
10. Clustering for Feature Engineering	10%	1 page

Submission

Your submission should be a `.zip` file containing (1) a PDF for your report and (2) a `.ipynb` **documented**, **reproducible Python notebook** as **rough work**. There is no need to submit the dataset, but it is expected that one is be able to reproduce your work on the copy of the dataset available on Moodle. You will find the submission link in our unit's Moodle area.

Please be careful **not to leave it to the last minute**; internet issues and similar do not count as exceptional factors.

Your report should contain your **name** and **student ID**. It **does not** need introduction, conclusion, or extra prose; simply focus on the structure shown above and add corresponding sections with content as suggested. You can create your report with any suitable text editor/processing system (e.g., Word, LaTeX); you can export plots/figures from Jupyter notebooks, for example, via `matplotlib`'s `.savefig()` functionality.

How to Pass This Assignment

One route to obtaining at least a pass in the assignment is to have a genuine, but limited attempt at 1-5, 7, 9.

Support Available

As always, you can ask questions regarding the coursework specification and obtain general feedback on your work in our scheduled, practical sessions, as well as at our office hours and at any extra sessions arranged. Our contact details are in the unit's Moodle area.

Regulations and Code of Conduct

Please make sure you are familiar with the Taught Postgraduate Assessment Regulations. As mentioned in the induction week, the pass mark is 50%, and one has a single reassessment opportunity (capped at 50%). If there are mitigating factors, please visit and follow the guidance at Exceptional Factors, such as that on self-certification.

Importantly, please also make sure you are fully aware of the regulations on Academic Misconduct, particularly if this is your first experience with the UK's higher-education system.

One important aspect to highlight is that this is an **individual assignment**, and the **submission has to be your own**. It is absolutely fine for people to work in groups and collaborate. It is also fine to **be inspired** by existing code snippets created by colleagues, contributors at StackOverflow, and LLMs such as **ChatGPT**. But, importantly, you have to **own it** in the end. That is, **you must be able to explain, customise, and apply it** in a similar, but separate context; also, **cite/reference** the sources. If in doubt, question yourself whether what the help you are receiving is "advice" or simply "do-it-for-me"; or whether you are simply "copy-and-pasting" instead of "owning" a code snipped.

Some examples of scenarios of when things **are not fine** are:

- *the representation of another person's work, without acknowledgement of the source, as one's own*: Say, student A wishes to help and shares their solution to the assignment with student B. The latter submits what was shared as their own work (fully or partially, identically or with little modification). This characterises **collusion** and would implicate both A and B.
- *the use of third parties and/or websites to attempt to buy assessments or answers to questions set*: this characterises **contract cheating**.

All cases of Academic Misconduct (e.g., plagiarism) will be reported to and investigated by Student Case Management Team.

Vivas

We will hold vivas (i.e., presentations with Q/A) for selected submissions - for example, when we need further clarification on the work described, when we have not been able to fully align the rough work with the report, when students fail to be in the labs carrying out the provisional work, or when we simply need to establish that the work submitted is the student's own. The performance on the viva will be used to confirm or adjust the provisional marks obtained on the submitted work.

Some Efficient and Effective Work Practices

- For **reducing computational time** and **cognitive load**, specially during prototyping/experimental phases, you might want to:
  - Employ packages such as `fireducks` and `cudf` to make your pandas much faster.
  - Avoid plotting a large number of individual data points; you could turn to heatmaps/hexbins instead of scatterplots, for example.

  - Take/work with a smaller sample of the dataset.

- When dealing with high-cardinality categorical features, at initial stages, you might want to reduce the number of categories before encoding, plotting, among others (say, keep the most frequent ones).

- An iterative approach with gradual expansion of scope and complexity is always recommended. For example, you might restrict your analysis initially to subsets of the data to features and regions of the data that are easier to make sense of.

- Favour things that can be automated and have a greater impact in the quality of the work. We wouldn't like to see you spending an inordinate amount of time in data cleaning, for example.