Manchester Metropolitan University

Department of Computing and Mathematics

ASSIGNMENT COVER SHEET

Unit code and title:	6G7V0015 Machine Learning Concepts
Assignment set by:	Luciano Gerber
Assignment ID: *(e.g. 1CWK50)*	1CWK100
Assignment weighting:	100%
Assignment title:	Machine Learning Project
Type: (Group/Individual)	Individual
Hand-in deadline:	Fri, 20 Jan 23 21:00
Hand-in format and mechanism:	Moodle; PDF, Jupyter Notebook

Learning outcomes being assessed:

LO1: Critically analyse a given problem and build predictive models to gain insights from real-world data.

LO2: Report on predictive outcomes and critically assess strengths and drawbacks of various predictive modelling techniques.

LO3: 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. Do not alter your work after the deadline. You should make at least one full backup copy of your work.

Penalties for late hand-in: see Assessment Regulations for Postgraduate Programmes of Study on the Student Life web pages. The timeliness of submissions is strictly monitored and enforced.

All coursework has a late submission window of 7 days (i.e., 5 working 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 below for further information on extensions.

Please note that individual tutors are unable to grant extensions to coursework.

Extensions: For most coursework assessments, you can request a 7-day extension through Moodle. This will not apply to in-class tests, presentations, interviews, etc., that take place at a specific time. If you need a longer extension you can apply for an Evidenced Extension. For an Evidenced Extension you **MUST**

submit 3rd party evidence of the condition or situation which has negatively impacted on your ability to submit or perform in an assessment.

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 Handbook (http://www.mmu.ac.uk/academic/casqe/regulations/docs/policies_regulations.pdf and Regulations for Postgraduate Programmes (https://www.mmu.ac.uk/academic/casqe/regulations/assessment/docs/pg-regs.pdf). Bad referencing or submitting the wrong assignment may still be treated as plagiarism. If in doubt, seek advice from your tutor.

If you are unable to upload your work to Moodle: If you have problems submitting your work through Moodle you can email it to the Assessment Team's Contingency Submission Inbox using the email address submit@mmu.ac.uk. You should say in your email which unit the work is for, and ideally provide the name of the Unit Leader. The Assessment team will then forward your work to the appropriate person. If you use this submission method, your work must be emailed by the published deadline, or it will be logged as a late submission. Alternatively, you can save your work to your university OneDrive and submit a Word document to Moodle which includes a link to the folder. It is your responsibility to make sure you share the OneDrive folder with the Unit Leader.

As part of a plagiarism check, you may be asked to attend a meeting with the Unit Leader, or another member of the unit 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.

Assessment Criteria:	Indicated in the attached assignment specification.
Formative Feedback:	Formative feedback on provisional work will be given primarily in lab sessions.
Summative Feedback Format:	Summative feedback is provided with a breakdown of marks for the assessment criteria and a sample solution.

(Draft) Coursework Specification (1CWK100)

6G7V0015 Machine Learning Concepts

Luciano Gerber

22/23, Semester 1

Overview

In this assignment students have the opportunity to consolidate their learning in the Machine Learning Concepts unit by creating a **working solution** for a **real-world, industry-based machine learning problem**. It involves typical tasks such as data **understanding**, **exploration**, **cleaning**, and **preparation**; and machine learning model **selection**, **tuning**, **fitting**, **evaluation**, and **analysis**.

More specifically, you will work with a **Car Sale Adverts dataset** provided by **AutoTrader (AT)**, one of our industry partners. The dataset contains an anonymised collection of adverts with information on vehicles such as brand, type, colour, mileage, as well as the selling price. Your main task is to produce a **regression** model for **predicting the selling price** given the characteristics of the cars in the historical data given. This type of task is similar to that addressed by AT's machine learning engineers and data scientists in the implementation of the price indicator feature in AT's website.

You are expected to produce, as **deliverable**, a **short, structured report** containing fragments of **code** and **explanations** for addressing a set of machine learning tasks on the dataset. In addition, you are asked to hand in a **fully-documented**, **reproducible Python notebook** with your **rough work**.

The **assessment criteria** of code/explanations involves aspects such as usefulness, correctness, clarity, conciseness, among others.

Obtaining and Exploring the Dataset

Please note the following **license information** before downloading and using the dataset:

- *this dataset provided by AutoTrader is to be used only by students enrolled on 6G7V0015 (22/23). Students can place it in their personal storage spaces (e.g., MMU OneDrive) for carrying out their practical data science and machine learning work (e.g., with Jupyter Lab). Please do not redistribute the dataset.*

It is available in our Moodle area here as a `.csv` file, and it has around 400K rows. For **reducing computational time** and **cognitive load**, specially during prototyping/experimental phases, you

might want:

- Take/work with a small **sample of the dataset**.
- Avoid plotting a large number of individual data points; you could turn to heatmaps and hexbins instead of scatterplots, for example.
- For high-cardinality categorical features: reduce the number of categories before encoding, plotting, among others.

Data Science Tasks to Perform

In order to address the main task of producing a regression model for price prediction, you need to implement and run the following tasks that are typical of a Machine Learning pipeline, using our usual environment (i.e., Python ecosystem/notebooks):

1. **Data/Domain Understanding and Exploration** (e.g., load the data, sample observations, check correct parsing of data, identify quantitative and qualitative features, analyse data distributions (e.g., range, centrality, dispersion, shape), identify good predictors, process the data for visualisation and exploration).
2. **Data Processing for Machine Learning** (e.g., detect and deal with noise (i.e., erroneous values), missing values, and outliers; subset, reshape, and engineer features for machine learning purposes; categorically-encode, rescale data; split data into predictors and target; obtain train/validation/test folds).
3. **Model Building** (e.g., choose suitable algorithm(s), fit and tune models; grid-search, rank, and select model(s) on based on evaluation metrics and under/overfit trade-off).
4. **Model Evaluation and Analysis** (e.g., evaluate selected model(s) according to popular score and loss metrics with cross-validation, analyse true vs predicted plot, gain and discuss insights based on feature importance, analyse individual predictions and distribution of scores/losses together with predictors).

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

Please structure your report with the components below including, for each, a small but representative and interesting portion of your work in completing the tasks above. These should consist of a relevant fragment of code, a corresponding output (e.g., table, plot), and a brief explanation (one

or two paragraphs) of findings. Next to each subcomponent below you will see a range (e.g., for *Dealing with Missing Values, Outliers, and Noise*, it is *1-2*) which indicates the suggest number of chosen examples to report on.

Component	Weight
1. **Data/Domain Understanding and Exploration**	**30%**
1.1. Meaning and Type of Features; Analysis of Univariate Distributions (3-4)	10%
1.2. Analysis of Predictive Power of Features (2-3)	10%
1.3. Data Processing for Data Exploration and Visualisation (1-2)	10%
2. **Data Processing for Machine Learning**	**20%**
2.1. Dealing with Missing Values, Outliers, and Noise (1-2)	10%
2.2. Feature Engineering, Data Transformations, Feature Selection (2-3)	10%
3. **Model Building**	**20%**
3.1. Algorithm Selection, Model Instantiation and Configuration (1-2)	10%
3.2. Grid Search, and Model Ranking and Selection	10%
4. **Model Evaluation and Analysis**	**30%**
4.1. Coarse-Grained Evaluation/Analysis (1-2) (e.g., with model scores)	10%
4.2. Feature Importance (2-4)	10%
4.3. Fine-Grained Evaluation (1-2) (e.g., with instance-level errors)	10%

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 are unlikely to 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.

How to Pass This Assignment

One pathway (narrow/deep) to obtaining at least a pass in the assignment would be to focus on and do a reasonable job on components 1 and 2 and a genuine start at 3. Say that you obtain 72% for (1), 72% for (2), 52% for (3), and 22% for (4) - that would result in a combined mark of 53%.

Another (broad/shallow) is to have a genuine, but limited attempt at each component. One possible scenario would be 62% for (1), 62% for (2), 52% for (3), and 52% for (4) - that would give you 57%.

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; in fact, this is something that I encourage. It is also fine to be inspired by existing code snippets created by colleagues, contributors at StackOverflow, and the like. 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.

A couple 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.