Queen Mary
University of London

MTH786P Machine Learning with Python, Semester A, 2023/24

Final project

1^{st} examiner Dr. N. Perra, 2^{nd} examiner Dr. O. Bobrowski

---

For the final project you will have to:

1. Select **one** of the five datasets provided on the QM+ module page and described below;
2. Analyse the dataset by studying the statistical characteristics of features and target variables. Plot and discuss their frequency/distributions and correlations. You are allowed to use Pandas, seaborn, and other libraries for plotting purposes;
3. Apply the relevant algorithms discussed during the module to solve a prediction task;
4. Write a detailed report (written preferably in LaTeX) of no more than 8 pages length structured as described below;
5. Submit a zipped folder with: i) the notebook you wrote with all the codes and results, ii) the data needed to run the codes. The name of the folder should be your student ID.

While, as mentioned above, you are allowed to use different libraries for visualisation purposes (e.g., Seaborn and Pandas), you can only use NumPy to implement your algorithms/models; tools from libraries such as SciPy/Scikit-learn, etc. are not allowed.

Datasets description (select one)

You will need to select **one** of the following datasets:

Credit card approval: credit card dataset

If you select this dataset, the task is to implement, describe and present (binary) classification models of your choice to predict which credit card applications are likely to be approved/rejected. More precisely, the goal is to implement (binary) classifiers using any number of the variables provided, which describe applicants' features, to predict the outcome of their credit card application. The dataset will be provided on the QM+ module page.

Disease prediction: Diabetes dataset

If you select this dataset, the task is to implement, describe and present (binary) regression/classification models of your choice to assess which characteristics are associated with diabetes. The goal is to develop (binary) regression/classification models using any number of the variables provided, which describe patients' features, to assess whether they are affected by diabetes or not. The dataset will be provided on the QM+ module page.

Wine quality prediction: White Wine dataset

If you select this dataset, the task is to implement, describe and present regression and/or classification models of your choice to predict the quality of white wines given a range of their features. The goal is to develop regression and/or classification models using any number of the variables provided, which describe wines' features, to predict their quality (measured as a score from 0 to 10 based on sensory data from three experts). The dataset will be provided on the QM+ module page.

Customers churn prediction: Bank dataset

If you select this dataset, the task is to implement, describe and present (binary) regression/classification models of your choice to predict which customers will churn (i.e., leave the bank). The goal is to develop (binary) regression/classification models using any number of the variables provided, which describe customers' features, to predict their churn. The dataset will be provided on the QM+ module page.

Usage prediction: Bikes dataset

If you select this dataset, the task is to implement, describe and present regression/classification models of your choice to predict the hourly number of bike rentals. The goal is to develop regression/classification models using any number of the variables provided, which describe weather's features, to predict the number of bikes rented. The dataset will be provided on the QM+ module page.

Structure of the written report

You conclude the assessment by writing a report. The report should be no longer than eight pages and be written preferably in LaTeXwith your favourite editor. If no editor is at hand, please feel free to use online editors such as Overleaf. The structure of the written report should be the following:

1. Title;
2. Your name and ID;
3. Introduction and problem statement;

4. Analysis of the dataset. Dive into the features and targets. Study their key statistical characteristics by plotting and discussing their frequency/distributions and correlations;

5. Methods. Describe in some details the algorithms/models you will be using. Use the notation and definitions discussed in the lectures and lecture notes;

6. Results of the prediction task. Present your results, compare the performance of the models implemented, and discuss the best you found;

7. Conclusions. Provide a summary of the problem you tackled and reflect on the results you found.

8. References. Reference i) the data you use, ii) the external resources like websites, articles, you might find and use, iii) the methods and approaches used. These do not count against the 8 page limits.

You need to submit also a zipped folder, via a dedicated submission link, with: i) the notebook you wrote with all the codes and results, ii) the data needed to run the codes (name the folder as your student ID). This is key to show the work done. Failure to submit the code will affect the final mark.

We refer to the following table for marking guidance on the final project. Note how the choice of dataset does not impact your final grade.

Approximate weight	**Description**
30 %	Overall quality of the project work (as described in the report), taking into account its difficulty.
30 %	Quality/relevance of the methodology and choice of resources, the application of these resources, explanation/exposition and correctness of mathematical notation and computer programming and analysis of results.
30 %	Quality of presentation: Logical structure of the document, clarity and coherence of exposition, correct use of English (punctuation, spelling and grammar) and precise mathematical writing, layout and style, sensible use of sections and subsections, lack of typing mistakes, choice of sensible (and standard) notation and other conventions, sensible use of equation numbering, appropriate use of figures, tables, charts, diagrams, etc.
10 %	Initiative and ambition (as described in the report): How far did you go, or attempt to go? How many models did you try/implement?