Queen Mary
University of London

ECN384 2024/25
Final Written Assessment Instructions

Submission information
The final written assessment is worth 70% of the overall module mark, and must be uploaded by **5pm on Wednesday 8th January 2025** in the designated submission area on the module's QMplus page in the 'assessment' tab. Please submit ONE file, in PDF or Word format (3,000 words) and the Jupyter Notebook file with your analysis.

You are advised to upload your file well in advance of the deadline, to avoid a late submission penalty resulting from technical problems (with QMplus availability, internet connectivity, etc.). For any technical issues, please contact ugsefsupport@qmul.ac.uk

Late submission penalty
The Queen Mary late submission policy is as follows: For every period of 24 hours, or part thereof, that an assignment is overdue there shall be a deduction of 5 marks. After seven calendar days (168 hours or more late) the mark shall be reduced to zero.

If you submit work late due to circumstances that are unforeseen and beyond your control, you should submit an extenuating circumstance claim on MySIS. If accepted, your work will be marked as normal with no late penalty.

Extensions
If you are unable to meet the submission deadline, due to circumstances that are unforeseen and beyond your control, then you may submit an extenuating circumstance claim on MySIS to request an extension. Please note that extensions can be granted for a maximum of one week.

Any EC claim that necessitates a deadline extension of more than one week will result in a new submission date being set in the Late Summer Examination Period.

Word count
If a response exceeds the indicated word count, then the examiners are not obligated to read or to mark the excess material. **The maximum number of words is 3,000.**

There is no penalty for a submission being below the word count, unless its brevity leads to substantive problems with the work.

Plagiarism and referencing

Please be aware that your work will be processed through the anti-plagiarism software called Turnitin. Please ensure that the work you submit is your own and that you have read the sections on referencing and on preparation of written work in the UG handbook.

Using AI and Large Language Models

Please see the guidance for students on Large Language Models (such as ChatPGT).

You can use AI tools and Large Language Models (like ChatGPT) to help you plan, write, and edit your assignment. 'Help' means that you should modify the output and should not submit it word-for-word.

Just like any other Internet source, you must

- Appropriately reference the origin of any LLM-generated text or ideas incorporated into your work, following the same referencing conventions as other sources.
- Save your original prompts and the machine-generated output (for example, using screenshots) and include this documentation in an Appendix that you submit along with your work.

The assignment

Data and objectives

LendingClub is an American peer-to-peer lending company, currently the world's largest platform that allows for individuals to both invest and borrow on the platform. Borrowers can obtain unsecured personal loans from the platform, and this coursework is set up for you to assess your ability to predict defaulters in the data using the predictors provided in the data. The data is a random sample of loans issued on the platform between 2007 and 2015, including the "loan_status" (the target variable) and payment information (predictors).

Coursework Instructions

You should submit one file, in PDF or Word format and a Jupyter Notebook file with your analysis. You will be provided with a .ipynb file (a Jupyter notebook) with specific instructions and with a .csv file that should be used for the assignment. I should be able to run the code on my computer by loading the same data file without any error. The problem you are investigating is to predict the defaulted loans using 1) the logistic regression model without cross – validation, 2) the classification tree model without cross – validation, 3) the classification tree model with cross – validation, and 4) the random forest.

For the report you should follow these sections:

- **Section one: Introduction**

Give a brief introduction on the problem you are investigating, and the models you will consider tackling this problem. Give a general "big picture" of your findings and your conclusion.

- **Section two: Methodologies**

Give brief explanation of the methods you used for data cleaning, preliminary analysis, as well as models you employed to analyse the dataset. Briefly explain the underlying theories behind the model you are employing. Point out pros and cons of each model.

- **Section three: Main findings**

In this section you should present the data pre - processing analysis and your estimation results for each model in relation to the data and model you are employing. Compare models/results and draw your conclusion. When comparing model performance across various methods, offer intuitive reasoning concerning the model's assumptions, advantages, disadvantages, and compatibility with the dataset, where applicable. Include all your figures and tables you may have obtained here.

- **Section four: Conclusion**

Summarise the problem you are investigating, draw your concluding remarks based on your findings.

Marking criteria

- Part A (40%)

    1. Set - up, visualize and pre - process the data (20%).

    2. Logistic regression and classification tree **without** cross validation (10%).

    3. Classification tree **with** cross validation and random forest (10%).

- Part B (60%)

    In this part, you are required to produce an extended report based on the work on part A. You should give an introduction on the problem you are analysing as well as the models you will consider tackling this problem. You should present the data analysis and your results for each model. Finally, you should compare the models/results and draw your conclusion.