Course Project Guidelines

Empirical Finance & Machine Learning

1 Project Overview

The goal of this course project is to apply machine learning techniques to the domain of empirical asset pricing using the provided high-dimensional dataset.

While the standard task is to *predict monthly stock returns*, you are free to define any reasonable task relevant to the data (e.g., *volatility prediction, regime classification, or portfolio optimization*).

2 Data Selection Policy

Students may choose between different data configurations.

- *Option A: The Full Sample*— Range of Years 1957 – 2021 (Full History).
- *Option B: Recent History (Lightweight)*
  - *Range: 2001 – 2021* (or a smaller subset if needed).
  - Faster training; easier debugging; fits easily within standard RAM limits.
  - You should justify your sample selection in the presentation (e.g., focusing on the modern algorithmic trading era, or reducing data size for computational feasibility).

3 Group Formation

- *Team Size:* You may work individually or in groups of 2.
- *Formation:* Students who intend to work in groups are expected to form groups on their own. Please finalize your grouping before the speed presentation.

4 Timeline and Milestones

The project is divided into two phases to ensure continuous progress and feedback.

Week 6: Speed Presentation (Formative)

- *Objective:* A "lightning talk" to present your proposal and initial data exploration.
- *Time Limit: 3 minutes presentation + 1 minute Q&A/Feedback.*
- *Content:* e.g., your task definition, data choice, validation scheme, and baseline model.
- *Grading:* Not graded (Feedback only).

Week 7: Final Presentation

- *Objective:* Present your final methodology, model performance, and insights.
- *Time Limit: 8 minutes presentation.*
- *Grading:* Graded (Evaluation based on Instructor + Peer Review).

5 Evaluation Guidelines

The project evaluation will draw inspiration from top-tier machine learning conference review guidelines (e.g., *ICML, NeurIPS*). A significant portion of the grade will be derived from *Peer Review.* When preparing your presentation, consider the following criteria:

1. ***Innovation & Design Rationale***
   - We encourage you to design your own architectures, e.g., custom Neural Network structures, hybrid models, rather than relying solely on off-the-shelf implementations.
   - Justify your design: explain *why* you designed the model for this task, e.g., whether or not to utilize the time series structure.
2. ***Understanding***
   - Rather than aiming at “accuracy”, we value deep understanding, e.g., whether and why the model performs in this way.
   - Does it demonstrate a clear grasp of the chosen model’s assumptions and limitations?
   - Is the entire workflow consistent and scientifically rigorous? e.g., is the validation strategy (such as a rolling window) strictly aligned with the time-series nature of the data to prevent look-ahead bias?
3. ***Clarity and Significance***
   - Is the presentation well-organized and easy to follow? e.g., raw code snippets may not appear very informative.
   - Are the presented results effective in conveying key insights? e.g., using cumulative return plots to compare strategies may be more convincing than reporting MSES.
   - Is the narrative compelling? Are the conclusions supported by the evidence?

6 Submission Deliverables

In the submission of final project, each group must submit:

1. *Presentation Slides.*
2. *Codebase:* A clean `.ipynb` (Jupyter Notebook) or coding script.

7 AI Policy & Integrity

The use of AI coding assistants (e.g., ChatGPT, GitHub Copilot) is *permitted.* You must include an *AI Acknowledgment* section in your slides stating *which tools were utilized for what purpose.*