# CQF Exam Three

Machine Learning

Jan 2025 Cohort

Instructions: The submitted report must present work and outputs clearly separated by Question. Submit ONLY ONE zip file named LASTNAME.zip that includes pdf file, code, html, data and any other supporting or working files. Python notebook with auxiliary output (data, plots) is not an analytical report: such submission will receive a deduction.

Please do not discuss this assignment in groups or messengers. Raise a support ticket for your queries. Only clarifying questions are allowed.

Introduction: Short-term asset returns are challenging to predict. Efficient markets produce near-normal daily returns with no significant correlation between $r_t$, $r_{t-1}$. This exam is a limited exercise in supervised learning. You are expected to explore multiple features of your choice, with both the original and final selected features being sufficiently numerous.

### Objective

Your objective is to develop a model to predict positive market moves (uptrend) using machine learning techniques as outlined in the section below. Your proposed solution should be comprehensive, including detailed feature engineering and model architecture.

- Choose one ticker of your choice from the index, equity, ETF, crypto token, or commodity.
- Predict the trend for short-term returns using binomial classification. The dependent variable should be labeled as [0, 1], not [-1, 1].
- The analysis should be comprehensive, including detailed data preprocessing, feature engineering, model building, tuning, and evaluation.

  Devise your own approach for categorizing extremely small near-zero returns (e.g., drop from the training sample or group with positive/negative returns). The threshold will depend on your chosen ticker. *Example:* small positive returns below 0.25% can be labeled as negative.

The number of features to include is a design choice, and there is no universally recommended set of features for all assets. The length of the dataset is also a design choice. For predicting short-term returns (e.g., daily moves), training and testing over a period of up to 5 years should be sufficient. Interpreting the instructions below is part of the task; the tutor will not assist in designing your computational implementation.

## A. Explanation of Entropy in Classification [10 marks]

**1.** What does entropy reveal about the quality of the partitions in a classification problem?

Answer below with True / False and explain the reasoning behind your choice.

(a) High entropy means the partitions are pure.

(b) High entropy means the partitions are impure."

## B. Feature Selection Using the Funnelling Approach [20 marks]

**2.** Perform feature selection for a machine learning model using a multi-step process by combining techniques from filter, wrapper, and embedded methods.

(a) Explain the feature selection process using the three categories of feature selection methods, step by step.

(b) Justify the selection of features retained at each step.

(c) Provide the final list of selected features.

## C. Model Building, Tuning and Evaluation [70 marks]

**3.** Predicting Positive Market Moves Using Gradient Boosting,

(a) Build a model to predict positive market moves (uptrend) using the feature subset derived above.

(b) Tune the hyperparameters of the estimator to obtain an optimal model.

(c) Evaluate the model's prediction quality using the area under the receiver operating characteristic (ROC) curve, confusion matrix, and classification report.

**Note**: The choice of boosting algorithm, and the number of hyperparameters to be optimized for the best model are design decisions. Simply presenting Python code without proper explanations will not be accepted. The report should present the study in detail, with a proper conclusion. As an optional add-on, consider backtesting the predicted signals as applied to a trading strategy.

* * *