CQF Exam Three

Machine Learning

Jan 2023 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. Address all portal and upload queries to CQFProgram@fitchlearning.com and clarifying only questions to Kannan.Singaravelu@fitchlearning.com.

Introduction: Short-term asset return is a challenging quantity 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: use a set of features from Table 1 without an expectation of predictive powers.

Objective

Your objective is to produce a model to predict positive moves (up trend) using machine learning models as specified in the below section. Your proposed solution should be comprehensive with the detailed feature engineering and model architecture.

- Choose one ticker of your interest from the index, equity, ETF, crypto token, or commodity.
- Predict trend only, for a short-term return (example: daily, 6 hours). Limit prediction to binomial classification: the dependent variable is best labeled [0, 1]. Avoid using [-1, 1] as class labels.
- Analysis should be comprehensive with detailed feature engineering, data pre-processing, model building, and evaluation.

  Devise your own approach on how to categorise extremely small near-zero returns (drop from training sample, group with positive/negative). The threshold will strongly depend on your ticker. *Example:* small positive returns below 0.25% can be labelled as negative.

Table 1: Features List

Feature	**Formula**	**Description**
O-C, H-L	Open - Close, High - Low	intraday price range
Sign	sign [rt = ln $\frac{P_t}{P_{t-1}}$]	sign of return or momentum
Past Returns	r_{t-1}, r_{t-2}, ...	lagged returns
Momentum	$P_t - P_{t-k}$	price change over k period
Moving Average	$SMA_i = \frac{1}{n}\sum_{i=0}^{n-1} P_{t-i}$	simple moving average
Exponential MA	$EMA_i = EMA_{t-1} + \alpha[P_t - EMA_{t-1}]$	recursive, $\alpha = 2/(N_{obs}+1)$

Number of features to include is a design choice. There is no one recommended set of features for all assets. Length of dataset is another design choice. If predicting short-term return sign (for daily move), then training and testing over up to 5-year period should be sufficient. Making sense of instructions below is part of the task: the tutor will not assist in designing your computational implementation.

A. Maths and Feature Engineering

1. Consider MSE($\hat{\beta}$) wrt to the true value β in context of regression methods,

$$E\left[(\hat{\beta}-\beta)^2\right] \quad = \quad Var[\hat{\beta}] \quad + \quad \Big(E[\hat{\beta}]-\beta\Big)^2$$

Answer below with Yes/No and one sentence of explanation referring to maths.

(a) can there exist an estimator with the smaller MSE than minimal least squares?

(b) for a prediction, does the MSE measure an irreducible error or model error?

2. What does entropy say about the partitions in a classification problem?

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

(a) high entropy means the partitions are pure.

(b) high entropy means the partitions are not pure.

3. Perform subset selection using any or all of a) filter, b) wrapper and, c) embedded methods

Note: Combination and selection of features set from above methods is a design choice.

B. Model Building and Evaluation

4. Using features selected from Part A,

(a) produce a model to predict positive moves (up trend) using support vector machine algorithm.

(b) tune hyperparameters for the estimator to present an optimal model.

(c) investigate model prediction quality using area under receiver operating characteristic curve, confusion matrix and classification report.

Note: Choice of kernels; method and number of hyperparameters to be optimized for the best model are design choices.

* * *