CQF Exam Three
Supervised Learning on Short-Term Asset Direction

2021 June Cohort

Instructions: Work on all questions is required (next page). Prepare a short report with headings that match Questions set in this exam. Submit ONE pdf file, named LASTNAME_REPORT_E3, and ONE zip file named LASTNAME_CODE.zip that includes code, data and any other files. Python notebook with code and auxiliary output (data, plots) is not a report: such submission will receive a deduction.

Please do not discuss this assignment in groups or messengers. Portal and upload questions to Orinta.Juknaite@fitchlearning.com. Clarifying only questions to Richard.Diamond@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.

- Choose **one ticker** of your interest form: equity, ETF, crypto token, or commodity.
  Do not choose: FX tickers (GBPUSD), equities with market cap over 100 bln. USD.
- Predict **direction only**, for a short-term return (daily, 6 hours). We limit prediction to binomial classification: dependent variable is best labelled $0, 1$ vs. $1, -1$.

  Devise own approach on how to categorise extremely small near-zero returns (drop from training sample or group with positive/negative).

Feature	Formula	Description
O-C, H-L	Open - Close, High - Low	of price
Sign	sign $[r_t = \ln \frac{P_t}{P_{t-1}}]$	sign of return, sign of momentum
Past Returns	$r_{t-1}, r_{t-2}, \ldots$	shift column of $t-1$ to obtain $t-2$
Momentum	$P_t - P_{t-k}$	price change period k days
Moving Average	$\text{SMA}_i = \frac{1}{n}\sum_{i=0}^{n-1} P_{t-i}$	simple moving average
Exponential MA	$\text{EMA}_t = \text{EMA}_{t-1} + \alpha\,[P_t - \text{EMA}_{t-1}]$	recursive, $\alpha = 2/(N_{obs}+1)$

Table 1: Features to choose from. Do not overlap, eg P_t and return for t.

There is no one recommended set of features for all assets. Making sense of instructions below is part of the task: the tutor will not assist in designing your computational implementation.

Length of dataset is another decision for you. If predicting short-term return sign (daily move), then training and testing over up to 5-year period should be sufficient. Train/test split and k-fold crossvalidation are optional – there is a little practical difference for daily moves direction prediction.

A. Feature Engineering and Penalised Classification

1. Identify a suitable set of features, generated from Table 1. Choose no less than 12 features initially.
   (a) this is your first practical task: a quick experiment on how many past returns to include $t-1, t-2, t-3, \ldots$. After, experiment to add Momentum of different length, and under 20-day SMA and/or EMA. Produce **a list** of features and their **correlation matrix**;
   (b) the question is down to quantity you are predicting: would it make sense to use 50D (fifty-day) Price Momentum to predict a one-day return (sign)?
   (c) decide on where to use scaling and **provide an explanation** why you think scaling is necessary or not necessary. Provide a table with the type of scaling/pre-processing for each feature.

2. Fit Ridge and Lasso logistic regressions and compare.
   (a) produce **a table** comparing L1 and L2 type of penalisation: the impact made on regression coefficients; plotting is optional here;
   (b) explain in bold font whether L1 or L2 regression likely to have a high bias and low variance;
   (c) plot **logistic sigmoid for three features** winning by largest coefficient and/or significance.

3. Return to a full set of features, implement feature scoring/elimination.
   Variance Inflation Factor • SelectKBest • Recursive Elimination • Shapley Values
   (a) choose at least two approaches and briefly indicate the main property and key maths for each. For example, VIF focuses on interdependent (colinear) features.
   (b) plot **the logistic sigmoid** for 3-4 winning features.

4. For the best model of your choice – restricted in features by Q2 penalisation or Q3 feature elimination (or combination of both approaches) produce **evaluation**: area under ROC curve plots for each class $0, 1$ and confusion matrix. Give expressions for precision/recall.

5. Call *predict_proba()* method. Provide **scatter plots of transition probabilities** of up moves, and another separate plot for down moves. It is required to use color-coding to indicate correctly/incorrectly predicted values on each scatter plot.

B. Mathematics of Supervised Learning

Question(s) in this section require mathematical working only. The tutor can't provide further hints.

6. Briefly present the maths of the logistic classifier by writing down: use β for coefficients not θ
   (a) the complete multivariate cost function;
   (b) log-loss form of the MLE log-likelihood function;

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

$$\mathbb{E}\left[(\hat{\beta}-\beta)^2\right] = \mathbb{V}\text{ar}[\hat{\beta}] + \left(\mathbb{E}[\hat{\beta}]-\beta\right)^2.$$

   Please 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?