









Modelling Financial Markets

Seminar 13: Machine Learning in Finance — Hands-on



Daniel Traian PELE
Antoaneta Amza

Bucharest University of Economic Studies
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Romanian Academy, Institute for Economic Forecasting
MSCA Digital Finance

Is a Bitcoin signal predictive under out-of-sample validation?

- Question of the seminar
  - **does any machine-learning signal on Bitcoin survive honest labels, honest validation and honest selection?**
- Deliverables
  - reproduce the solved A examples and solve A7 independently
  - one chart and one table per B exercise
  - the corrected C1 code
  - one interpretation sentence per exercise
- Notebook for today: open the seminar notebook in Colab
  - each task names its notebook section
- Seminar: Prof. Daniel Traian Pele, danpele@ase.ro; Antoaneta Amza, doctoral candidate, antoaneta.amza@csie.ase.ro

Returns and conventions across lecture and seminar

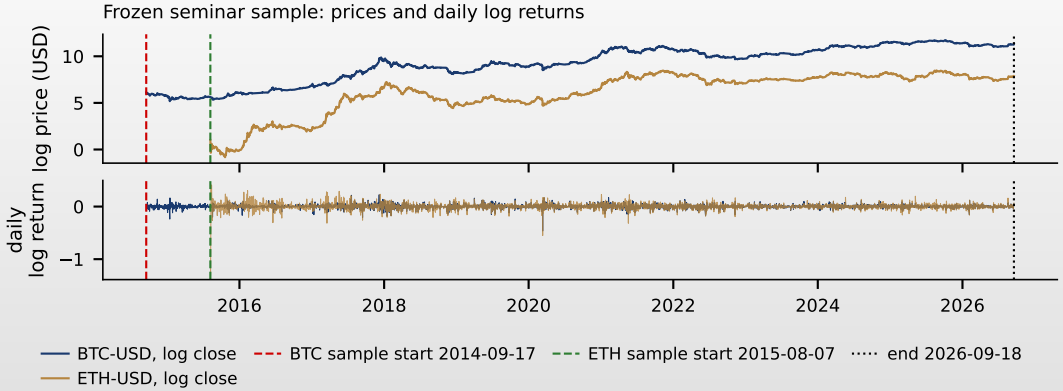
- Log returns in the notation are decimal quantities. A price rising from 100 to 110 has a 10% simple return and an approximately 9.53% log return. Multiply decimal returns by 100 to express them in percent.
- In EWMA (exponentially weighted moving average), α in the seminar weights new information and λ in the lecture weights the past: $\alpha + \lambda = 1$.
- For triple-barrier labels, pt and sl name the profit and stop-loss barrier multipliers, denoted π_u and π_l in the lecture.
- In Logit-L1 the penalty λ is the inverse of C ("The Logistic Model with an L1 Penalty"). The lecture plot uses an inverted $1/C$ axis; the seminar plot uses C: a higher penalty means a smaller C.

Data Used

Series	Field	Frozen sample	Used in
Bitcoin	close	2014-09-17 to 2026-09-18, 4,385 closes	B1, B3–B6, B8–B14, C1, C2
Ether	close	2015-08-07 to 2026-09-18, 4,061 closes	B7
VIX	close	US trading days, carried forward to the BTC calendar	B14
Kenneth French factors and deciles	Data Library, monthly	1967 – latest month	C3

- Daily closes of the course data, one series per asset (field used: close); crypto trades every day, so a year has $A = 365$ return periods
- Daily log returns (Notation) on each asset's own calendar; no missing closes in the frozen samples; events without a complete h-day window at the end are not labelled

Setup: Run and Verify (1/2)



Setup: Run and Verify (2/2)

```
# notebook section 0 defines load_data
btc = load_data('btc')['Close']
eth = load_data('eth')['Close']
print(len(btc), btc.index[0], btc.index[-1])
print(len(eth), eth.index[0], eth.index[-1])
r = np.log(btc).diff().dropna()
print(round(r.std() * np.sqrt(365), 3))
```

- Expected for BTC (Bitcoin): 4,385 closes, from 17 Sep 2014 (457.33) to 18 Sep 2026 (80,901.46); annualised volatility 0.666
- For ETH (Ether, the cryptocurrency of Ethereum): 4,061 closes, from 7 Aug 2015
- If a count differs, stop and check the series, the field and the dates

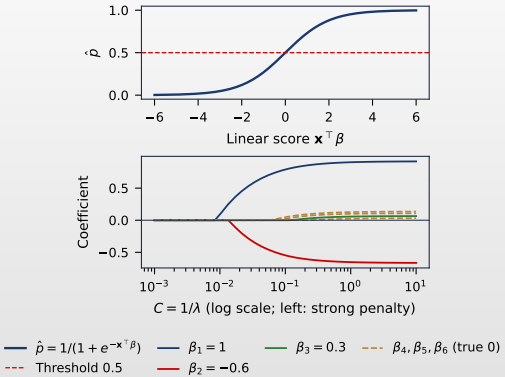
MFM_ch13_seminar

Notation and the Supplied Functions

- P_t: close on day t; $x_t = \ln P_t$; $r_t = x_t - x_{t-\mathbf{1}}$; B: backshift, $Bx_t = x_{t-\mathbf{1}}$
- Decision time: the close of day t; in the rule grid (B6, B12, B13) the trade is idealised at the same close, so the position earns $r_{t+\mathbf{1}}$; with execution at the next close it would earn $r_{t+\mathbf{2}}$
- d: order of a fractional filter, a choice ("Fractional Differencing"); δ: memory of a series, $I(\delta)$, an unknown parameter we estimate ("Memory, the ADF Test and Local Whittle")
- $\sigma_{t_{\mathbf{0}}}$: EWMA volatility of daily log returns, span 50 ("Fixed-Horizon Labels and EWMA Volatility"); h: label horizon; $[t_{\mathbf{0}}, t_{\mathbf{1}}]$: the span of a label
- SR: per-day Sharpe ratio of daily log returns, risk-free rate 0 ("Fixed-Horizon Labels and EWMA Volatility"); $\gamma_{\mathbf{3}}$, $\gamma_{\mathbf{4}}$: skewness and ordinary kurtosis ("The Probabilistic Sharpe Ratio")

Supplied function (notebook section 0)	What it returns
`ffd_weights(d, threshold)`, `frac_diff_ffd`	weights $w_{\mathbf{0}}, \ldots, w_K$; the filtered series from day K on
`local_whittle`, `exact_local_whittle`	$(\hat{\delta}, \mathrm{SE})$ with $m = \lfloor n^{\mathbf{0.65}} \rfloor$ frequencies
`get_daily_vol`, `triple_barrier`	EWMA volatility; one row per event: $t_{\mathbf{1}}$, return, label, exit type
`PurgedKFold`, `build_features`	purged and embargoed folds; the 11 features of B3
`probabilistic_sharpe_ratio`, `deflated_sharpe_ratio`	PSR (Probabilistic Sharpe Ratio) and DSR (Deflated Sharpe Ratio) of a per-day SR

The Logistic Model with an L1 Penalty



- Set-up: features x_t known at the close of day t, label $y_t \in \{0, 1\}$ ("up" = 1)
  - a classifier returns $\hat{p}_t = \hat{P}(y_t = 1 \mid x_t)$ and predicts "up" when $\hat{p}_t > 0.5$
- **Logit-L1** (Tibshirani (1996)): logistic link, coefficients shrunk by an absolute-value penalty

$$\hat{p} = \frac{1}{1 + e^{-\mathbf{x}^\top \boldsymbol{\beta}}}, \qquad \min_{\boldsymbol{\beta}} \sum_t \ell_t(\boldsymbol{\beta}) + \lambda \sum_j |\beta_j|$$

  - $\mathbf{x}^\top \boldsymbol{\beta} = \sum_j \beta_j x_j$: linear score; $\beta_j > 0$: a higher x_j raises $\hat{p}$
  - ℓ_t: log loss ("The Log Loss"); $\lambda \geq 0$: penalty weight; in scikit-learn $C = 1/\lambda$
  - lower panel: as C falls, weak coefficients become exactly 0

Trees, Random Forests and Boosting

- **Classification tree**: splits the feature space by rules $x_j \le c$; in leaf R it predicts the share of "up" labels

$$\hat{p}_R = \frac{1}{n_R} \sum_{t \in R} y_t, \qquad G(R) = 2\hat{p}_R(1 - \hat{p}_R)$$

  - n_R: observations in R; G: Gini impurity, 0 for a pure leaf; each split is chosen to reduce $\sum n_R G(R)$ most
  - minimum leaf size (e.g. 50): a larger value gives smoother, less overfit trees
- **Random forest** (RF, Breiman (2001)): average of B trees, each grown on a bootstrap sample, with a random subset of features at every split

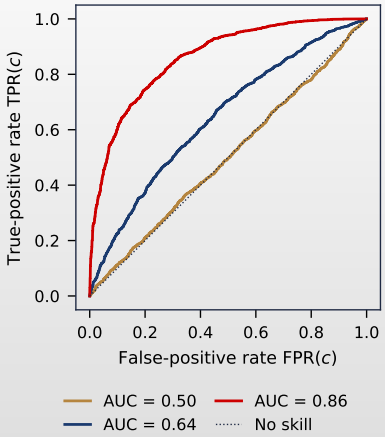
$$\hat{p}_{RF}(\mathrm{x}) = \frac{1}{B} \sum_{b=\mathbf{1}}^{B} \hat{p}_b(\mathrm{x})$$

- **Gradient boosting** (GB, Friedman (2001)): trees added one at a time on the log-odds scale, $\ln[\hat{p}/(1-\hat{p})]$

$$F_m(\mathrm{x}) = F_{m-\mathbf{1}}(\mathrm{x}) + \eta\, h_m(\mathrm{x}), \qquad \hat{p} = \frac{1}{1 + e^{-F_M(\mathbf{x})}}$$

  - h_m: a small tree fitted to the residuals $y_t - \hat{p}_{m-\mathbf{1},t}$ (the negative gradient of the log loss)
  - $\eta \in (0, 1]$: learning rate; M: number of trees; a small η learns slowly and overfits less

Accuracy, the ROC Curve and the AUC



- **Accuracy** $= \frac{1}{n}\sum_t 1\{\hat{y}_t = y_t\}$, $\hat{y}_t = 1\{\hat{p}_t > 0.5\}$
  - benchmark: "always up", whose accuracy is the share of up labels, not 50%
- **ROC curve** (receiver operating characteristic): for every threshold c, the true positive rate (TPR) and the false positive rate (FPR)

$$\mathrm{TPR}(c) = P(\hat{p}_t > c \mid y_t = 1), \quad \mathrm{FPR}(c) = P(\hat{p}_t > c \mid y_t = 0)$$

- **AUC** (Area Under the, ROC Curve) (Hanley & McNeil (1982)): area under the ROC curve

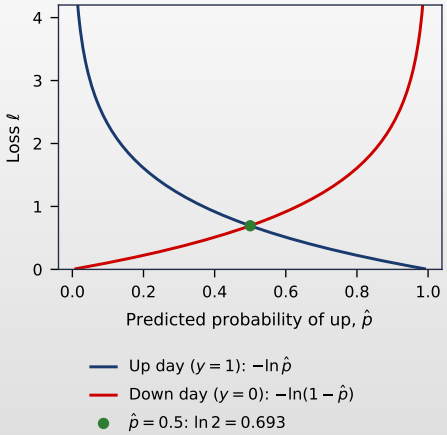
$$\mathrm{AUC} = P(\hat{p}_+ > \hat{p}_-) + \tfrac{1}{2}P(\hat{p}_+ = \hat{p}_-)$$

  - $\hat{p}_+$, $\hat{p}_-$: scores of a random up day and a random down day
  - 0.5: no ranking ability (diagonal); 1: perfect ranking; below 0.5: inverted
  - Example: up days $(0.7, 0.4)$, down days $(0.5, 0.2)$: 3 of 4 pairs ordered correctly, AUC $= 0.75$

Accuracy Against the Baseline

- Question for the room: does a 55% accuracy at 5 days on Bitcoin prove skill?
- Answer: no; 54.8% of the 5-day windows are "up" (B3a)
- The "always up" rule therefore already scores 54.8%; a model is compared with this baseline, not with 50%

The Log Loss



- **Log loss**: penalises confident wrong probabilities

$$\ell_t = -\big[y_t \ln \hat{p}_t + (1 - y_t) \ln(1 - \hat{p}_t)\big]$$

  - $\ell_t \geq 0$: $-\ln \hat{p}_t$ on an up day, $-\ln(1 - \hat{p}_t)$ on a down day
  - the constant $\hat{p}_t = 0.5$ scores $\ln 2 = 0.693$ every day
  - Example: $y_t = 1$; $\hat{p}_t = 0.6$ costs 0.51, $\hat{p}_t = 0.2$ costs 1.61
- **Brier score**: the squared error of a probability, $b_t = (y_t - \hat{p}_t)^2 \in [0, 1]$
- AUC judges the ranking; log loss and Brier judge the quality of the probabilities, calibration included: a model can rank well and still have a higher log loss (B4)

Comparing Two Forecasts

- Out-of-fold prediction: every day is predicted by a model that was not trained on it ("Validation without Information Leakage" says which days must also be left out)
- Comparing two forecasts A and B: the Diebold–Mariano (DM) statistic (Diebold & Mariano (1995))

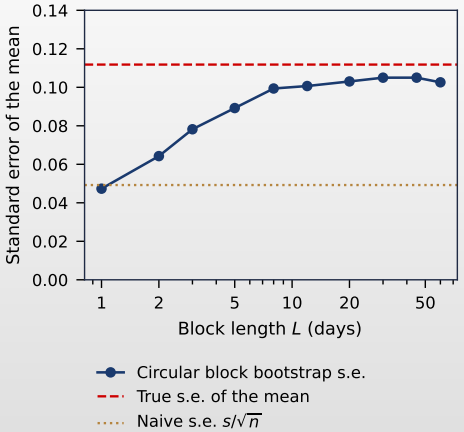
$$d_t = \ell_t^A - \ell_t^B, \qquad DM = \frac{\bar{d}}{\widehat{\mathrm{SE}}(\bar{d})}$$

  - ℓ_t^A, ℓ_t^B: losses of A and B on day t; $d_t > 0$: A worse that day; $\bar{d}$: mean of d_t
  - H_0: equal expected loss; DM is approximately standard Normal; $|DM| > 1.96$: a significant difference at 5%
- $\widehat{\mathrm{SE}}(\bar{d})$: Newey–West standard error (SE) (Newey & West (1987)), because neighbouring d_t are dependent

$$\widehat{\mathrm{SE}}(\bar{d}) = \sqrt{\frac{1}{n}\Big[\hat{\Gamma}_0 + 2\sum_{j=1}^{q}\Big(1 - \frac{j}{q+1}\Big)\hat{\Gamma}_j\Big]}$$

  - $\hat{\Gamma}_j$: sample autocovariance of d_t at lag j; q: number of lags (at least the label overlap)
  - equivalently: regress d_t on a constant with HAC (Heteroskedasticity and Autocorrelation Consistent) errors; the t-value is DM
  - the naive i.i.d. t uses only $\hat{\Gamma}_0$ and overstates significance when $\hat{\Gamma}_j > 0$

The Uncertainty of an AUC and the Block Bootstrap

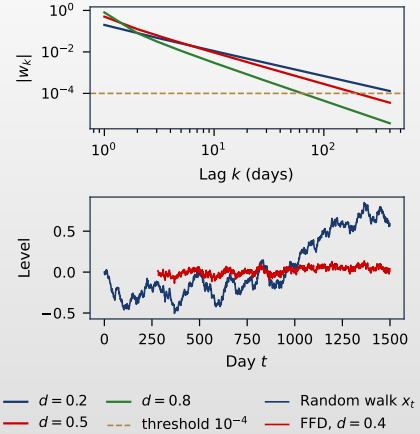


- **DeLong** (DeLong, DeLong & Clarke-Pearson (1988)): a closed-form SE of the AUC that treats the days as independent
  - $z = (\mathrm{AUC} - 0.5)/\mathrm{SE} \approx N(0,1)$ under H_0: AUC $= 0.5$
- **Circular block bootstrap** (Künsch (1989))
  - draw random starts, take blocks of L consecutive days (wrapping around the end) until n days
  - recompute the statistic $\hat{\theta}_b^*$ in each of B replications

$$\mathrm{SE}^* = \mathrm{sd}(\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*), \quad CI_{95} = [\hat{\theta}^*_{(2.5\%)},\ \hat{\theta}^*_{(97.5\%)}]$$

- CI_{95}: the 95% percentile confidence interval (CI)
- $L = 1$ is the ordinary bootstrap; blocks keep the dependence inside them
- Chart (simulated overlapping 5-day sums): $L = 1$ gives the naive SE; from $L \approx 10$ the SE is close to the true one

Fractional Differencing



- Differencing of order d (Hosking (1981)), with the backshift $Bx_t = x_{t-1}$ and the log price x_t

$$(1-B)^d x_t = \sum_{k \geq 0} w_k x_{t-k}, \quad w_0 = 1, \; w_k = -w_{k-1} \frac{d-k+1}{k}$$

  - k: lag in days; w_k: weight of the value k days back
  - $d = 0$ keeps the level, $d = 1$ gives the return $x_t - x_{t-1}$
  - $d = 0.5$: $w = (1, -0.5, -0.125, -0.0625, \ldots)$; slowly decaying weights keep part of the old levels: "memory"

- **FFD** (fixed-width window fractional differentiation, López de Prado (2018)): keep the weights until the first $|w_k| < \tau$ (e.g. 10^{-4}) and apply the same $K+1$ weights every day
  - K: last lag kept; a smaller d needs a longer window (upper panel)
  - lower panel: FFD at $d = 0.4$ of a random walk looks mean-reverting

Memory, the ADF Test and Local Whittle

- **Memory** δ: x_t is $I(\delta)$ when $(1-B)^{\delta} x_t$ has short memory
  - $\delta < \frac{1}{2}$: stationary; $\frac{1}{2} \leq \delta < 1$: non-stationary, mean-reverting; $\delta = 1$: unit root (random walk)
  - an untruncated filter of order d applied to an $I(1)$ series leaves memory $1-d$
- **ADF (Augmented Dickey–Fuller) test** (Dickey & Fuller (1979)): H_0: unit root ($\gamma = 0$) in the regression

$$\Delta x_t = c + \gamma x_{t-1} + \sum_{j=1}^{p} \phi_j \Delta x_{t-j} + \varepsilon_t$$

  - statistic: the t-value of $\hat{\gamma}$; Dickey–Fuller distribution, not t; reject if below -2.86 (5%, with a constant)
  - minimum-d rule (also called the minimum-ADF rule, López de Prado (2018)): the smallest d at which ADF rejects; A1 asks what that proves
- **Local Whittle** (Robinson (1995)): $\hat{\delta}$ minimises, over the m lowest Fourier frequencies $\lambda_j = 2\pi j/n$

$$R(\delta) = \ln\left(\frac{1}{m}\sum_{j=1}^{m} \lambda_j^{2\delta} I(\lambda_j)\right) - \frac{2\delta}{m}\sum_{j=1}^{m} \ln \lambda_j, \qquad \mathrm{SE}(\hat{\delta}) = \frac{1}{2\sqrt{m}}$$

  - $I(\lambda_j)$: periodogram; n: sample length; $m = \lfloor n^{0.65} \rfloor$, $\lfloor \cdot \rfloor$: integer part
  - exact local Whittle (Shimotsu & Phillips (2005)) is also valid for $\delta \geq \frac{1}{2}$

Fixed-Horizon Labels and EWMA Volatility

- **Fixed-horizon label** over h days

$$y_t = 1\{\ln P_{t+h} - \ln P_t > 0\}$$

  - $1\{A\} = 1$ if A holds and 0 otherwise; h: horizon in days
  - the label of day t uses the returns of $[t, t+h]$: neighbouring labels share returns

- **EWMA volatility** (exponentially weighted moving average) of daily log returns r_t

$$\sigma_t^2 = (1-\alpha)\,\sigma_{t-1}^2 + \alpha\, r_t^2, \qquad \alpha = \frac{2}{\text{span} + 1}$$

  - span 50: $\alpha = 2/51 \approx 0.039$ on the newest day; the weight of day $t-k$ is $\alpha(1-\alpha)^k$
  - `get_daily_vol` uses the EWMA standard deviation (demeaned); σ_{t_0} scales the barriers of the next slide

- **Sharpe ratio** (SR) of daily returns, risk-free rate 0

$$\widehat{SR} = \frac{\bar{r}}{\hat{\sigma}}, \qquad \widehat{SR}_{ann} = \widehat{SR}\sqrt{A}$$

  - $\bar{r}$, $\hat{\sigma}$: mean and standard deviation of daily returns; A: days per year (365 for Bitcoin, 252 for stocks)

The Triple Barrier

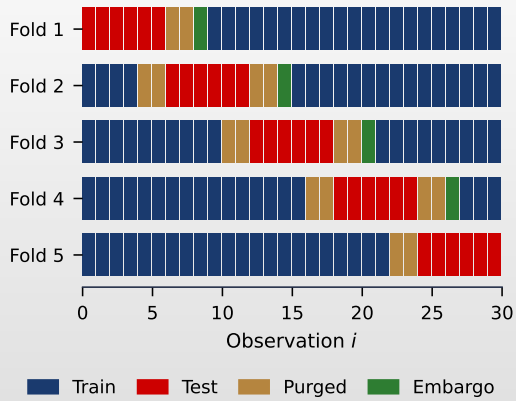


- **Triple barrier** (López de Prado (2018)): from the entry t_0, two horizontal barriers and a vertical one

$$P_{t_0} e^{+pt\,\sigma_{t_0}\sqrt{h}}, \qquad P_{t_0} e^{-sl\,\sigma_{t_0}\sqrt{h}}, \qquad t_0 + h$$

  - $\sigma_{t_0}\sqrt{h}$: typical h-day log move; pt, sl: profit-taking and stop-loss widths in units of it
  - label $+1$ / -1 at the first horizontal touch; otherwise the sign of the return at $t_0 + h$; t_1: exit day
- Chart (simulated): $\sigma = 2\%$, $h = 10$, $pt = sl = 1$: barriers 106.53 and 93.87; the close of day 4 is 107.02, so $t_1 = 4$ and the label is $+1$, although the path later falls
- Only daily closes are checked: a touch inside the day is missed

Validation without Information Leakage



- K-fold cross-validation (CV): train on $K-1$ folds, test on the remaining one
  - **shuffled** folds mix days; **contiguous** folds keep blocks of time
- **Purging** (López de Prado (2018)): drop from training every observation whose label span overlaps a test label span
- **Embargo**: also drop the e observations after the end of the last test label, because features look back
- Example: $n = 10$, spans $[i, i+2]$, test fold $\{4, 5\}$ (labels up to day 7), $e = 1$
  - drop $\{2, 3\}$ before, $\{6, 7\}$ after (they overlap the closed test span through day 7) and $\{8\}$ as the embargo; train on $\{0, 1, 9\}$
- Chart: $n = 30$, 5 folds, spans $[i, i+2]$, $e = 1$
- Without purging, a training day shares returns with a test day and a flexible model memorises the neighbour (B2)

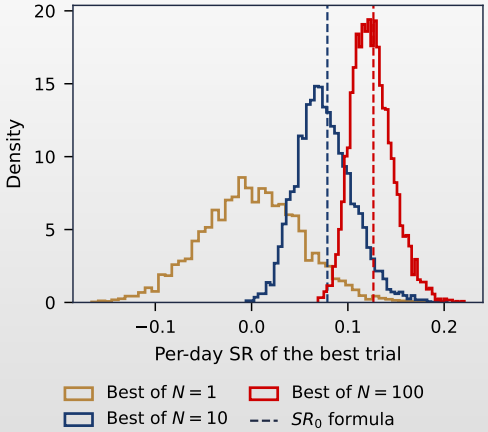
The Probabilistic Sharpe Ratio

- Probabilistic Sharpe ratio (PSR, Bailey & López de Prado (2012)): the Normal approximation to one minus the one-sided p-value for a per-day benchmark SR^*

$$PSR(SR^*) = \Phi\Big(\frac{(\widehat{SR} - SR^*)\sqrt{T-1}}{D}\Big), \qquad D = \sqrt{1 - \gamma_3 \widehat{SR} + \frac{\gamma_4 - 1}{4}\widehat{SR}^2}$$

  - $\widehat{SR}$: estimated per-day SR; T: number of returns; Φ: standard Normal distribution function
  - $\gamma_3 = \frac{1}{T}\sum_t (r_t - \bar{r})^3/\hat{\sigma}^3$: skewness; $\gamma_4 = \frac{1}{T}\sum_t (r_t - \bar{r})^4/\hat{\sigma}^4$: ordinary kurtosis (3 for the Normal distribution)
  - $D/\sqrt{T-1}$: standard error of $\widehat{SR}$; negative skewness and fat tails raise D and lower the PSR
  - $D = 1$ to first order for Normal returns ($\gamma_3 = 0$, $\gamma_4 = 3$, small $\widehat{SR}$)
- Reading: $PSR(0) = 0.95$ corresponds to a one-sided p-value of 5% against a nonpositive SR, under i.i.d. returns and the delta approximation, for **one** pre-specified strategy; it is not a posterior probability for the parameter

Selection and the Deflated Sharpe Ratio



- Selection: the best of N strategies with zero true skill still shows a positive SR; its expected value is

$$SR_{\mathbf{0}} = \sqrt{V}\left[(1-\gamma_E)\Phi^{-\mathbf{1}}(1-\tfrac{\mathbf{1}}{N}) + \gamma_E \Phi^{-\mathbf{1}}(1-\tfrac{\mathbf{1}}{Ne})\right]$$

  - V: variance scale used for trial Sharpe ratios; the null illustration uses $V = 1/T$, while empirical cross-trial dispersion is a different approximation; $\Phi^{-\mathbf{1}}$: standard Normal quantile function; $e \approx 2.718$; $\gamma_E = 0.5772$: Euler's constant

- **Deflated Sharpe ratio** (DSR, Bailey & López de Prado (2014))

$$DSR = PSR(SR_{\mathbf{0}})$$

  - the PSR against the luck of the best trial; above 0.95 gives a one-sided p-value below 5% against that adjusted benchmark, under the stated approximation

- Chart (simulated, $T = 400$, $V = 1/T$): the best SR moves right as N grows; dashed lines: $SR_{\mathbf{0}} = 0.079$ ($N = 10$) and 0.127 ($N = 100$)

Sharpe Ratios After Selecting the Best Strategy

- Per-day $\widehat{SR} = 0.10$, $T = 400$, $D = 1$: $PSR(0) = \Phi(0.10\sqrt{399}) = \Phi(2.00) = 0.977$
- If it is the best of $N = 10$ trials with $V = 1/T$
  - $\Phi^{-1}(0.9) = 1.282$, $\Phi^{-1}(1 - \frac{1}{10e}) = 1.789$, $\sqrt{V} = 1/20$
  - $SR_0 = \frac{1}{20}[0.4228 \cdot 1.282 + 0.5772 \cdot 1.789] = 0.079$
  - $DSR = \Phi\big((0.10 - 0.079)\sqrt{399}\big) = \Phi(0.43) = 0.66$
- Reading: alone, the SR is convincing (0.977); as the best of ten tries, it is not ($0.66 < 0.95$)
- More in Chapter 13: the false strategy theorem, feature importance, walk-forward backtests and the S&P 500 (Standard & Poor's) case study

Tools for the Extension Packs

- **Precision, recall, F1** of the decision "take the trade" (B11); TP, FP, FN: true positives, false positives, false negatives

$$\text{precision} = \frac{TP}{TP + FP}, \qquad \text{recall} = \frac{TP}{TP + FN}, \qquad F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- **Feature importance** (B3c)
  - MDI (mean decrease in impurity): the reduction of $\sum n_R G(R)$ due to splits on feature j, averaged over the trees; in sample
  - MDA (mean decrease in accuracy): $\mathrm{AUC} - \mathrm{AUC}_{(j)}$, with $\mathrm{AUC}_{(j)}$ the test AUC after randomly permuting feature j; measures the change in score on the evaluated set; read a negative value with the spread of the permutations and the estimation uncertainty
  - SHAP (SHapley Additive exPlanations, Lundberg & Lee (2017)): $\hat{p}(\mathsf{x}) = \phi_{\mathbf{0}} + \sum_j \phi_j(\mathsf{x})$, a fair split of one prediction among the features; it attributes the prediction, without the realised outcome; importance: mean $|\phi_j|$
- **Spearman rank correlation**: the Pearson correlation of the ranks; 1 = identical orderings

Exercise Map

Exercise	Question	Type	Model
A1	what does a truncated fractional filter keep of the price level?	Solved	–
A2, B1	how do triple-barrier labels behave on toy paths and on Bitcoin?	Solved	–
A3, B2	what does purging remove, and what does shuffling fake?	Solved	–
A7	how does a shorter label change the lost observations?	Proposed	A3
B3a, B3b	do three models beat "always up", and how uncertain is the AUC?	Solved	–
B4	does a model forecast better than a constant?	Proposed	B3a, B3b
C1	what is wrong in an AI answer?	Proposed	B2, B3a
A4	is the best of 100 backtests better than luck?	Solved	–
After the lecture	B5: does fractional differencing make Bitcoin stationary? B6: the best of 120 rules against luck	Solved	A1, A4
Extension packs	A2 Extended, A5, A6, A8–A10, B3c, B7–B14, C2, C3	both	on each task

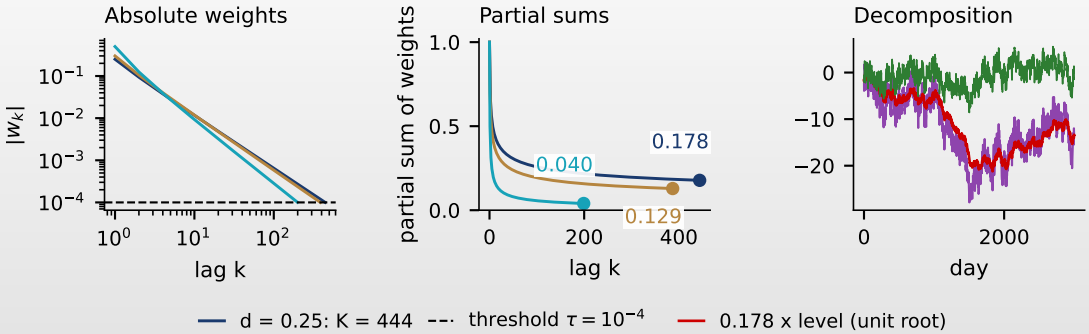
- **[Solved]**: full solution in the slides and in the notebook, a model to copy; **[Proposed]**: you solve it, the solution is discussed in class
- A: derivations and small numerical cases, on paper; B: inference on crypto data, in the notebook, with an interpretation question; C: an AI (Artificial Intelligence) critique and project seeds
- Part I: A1 to B2; Part II: B3a to A4; B5, B6 use lecture tools (exact local Whittle, effective number of trials)

A1 [Solved]: FFD Weights and the Retained Window — Task

- **Question**: what does a truncated fractional filter keep of a price level?
- **Inputs**: the FFD weights w_k and their truncation rule ("Fractional Differencing"), with $\tau = 10^{-4}$; K = the last retained lag, so the window uses $K+1$ observations; no data file
- **Tasks**
  1. Compute $w_0, \ldots, w_4$ for $d = 0.25$ with the recursion.
  2. With `ffd_weights(0.25)`, report K, the window length $K+1$ and $\sum_{k \le K} w_k$; repeat for $d = 0.3$ and $d = 0.5$.
  3. Write $\tilde{X}_t = \sum_{k=0}^{K} w_k X_{t-k}$ as $(\sum_k w_k) X_t$ minus a combination of the differences $X_t - X_{t-k}$.
  4. Decide whether $\tilde{X}_t$ of an $I(1)$ log price is stationary.
- **Report**: the five weights, a table (d, K, $K+1$, weight sum), the decomposition and the decision
- **Interpretation**: the minimum-d rule picks the smallest d at which an ADF test rejects: what does such a rejection prove?
- Notebook: section A1

Back Exercise list

A1 [Solved]: Solution (1/2) — FFD Weights and the Retained Window



A1 [Solved]: Solution (2/2) — FFD Weights and the Retained Window

1. $w_1 = -0.25$, $w_2 = -0.25 \cdot 0.75/2 = -0.09375$, $w_3 = -0.0547$, $w_4 = -0.0376$: every weight after w_0 is negative
2. Table:

d	K	$K+1$	$\sum_{k \le K} w_k$
0.25	444	445	0.178
0.30	387	388	0.129
0.50	199	200	0.040

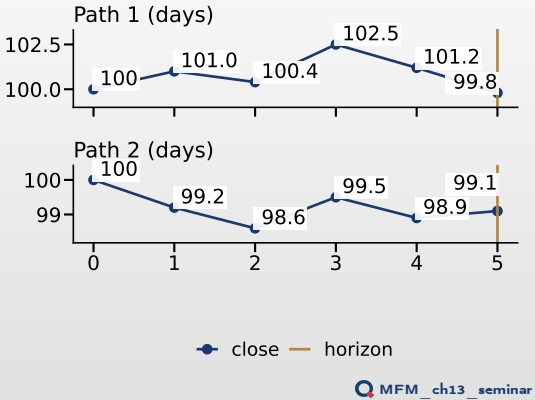
3. $\tilde{X}_t = (\sum_{k \le K} w_k) X_t - \sum_{k=1}^{K} w_k (X_t - X_{t-k})$
4. Each $X_t - X_{t-k}$ is $I(0)$, but the first term is $0.178\, X_t$ at $d = 0.25$: $\tilde{X}_t$ is still $I(1)$

- Right panel: one simulated random walk; the FFD series tracks the scaled level (red), which carries the unit root (correlation 0.93)
- A simple case: weights $(1, -0.5)$ give $X_t - 0.5X_{t-1} = 0.5X_t + 0.5(X_t - X_{t-1})$; one term keeps the nonstationary level, the other is a stationary increment
- Interpretation: in this known filtered $I(1)$ model, an ADF rejection is misleading; the algebraic decomposition, not the rejection, establishes nonstationarity

MFM_ch13_seminar

A2 [Solved]: Triple-Barrier Labels of Two Paths — Task

- **Questions**
  - which barrier is touched first?
  - which label results?
- **Inputs**: entry $P_{t_0} = 100$, daily volatility $\sigma_{t_0} = 1\%$, vertical barrier $h = 5$ days, $pt = sl = 1$; barriers as in "The Triple Barrier"; daily closes only
- **Tasks**
  1. Compute the upper and lower barrier.
  2. Draw both on the two paths and mark the first touch.
  3. Label each path: $+1$ upper, -1 lower, and at the vertical barrier the sign of $\ln(P_{t_0+h}/P_{t_0})$.
- **Report**: the two barriers, the exit day and type, the two labels
- **Interpretation**: why can a path that ends below 100 still get $+1$? Notebook: section A2



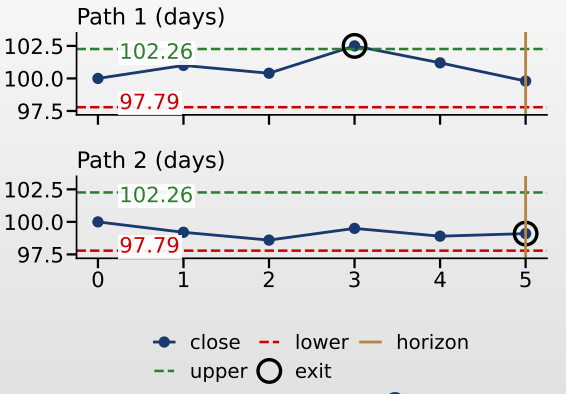
MFM_ch13_seminar

Back Exercise list

A2 [Solved]: Solution — Triple-Barrier Labels of Two Paths

1. Width $0.01\sqrt{5} = 0.02236$: upper $100e^{\mathbf{0.02236}} = 102.26$, lower $100e^{-\mathbf{0.02236}} = 97.79$
2. Path 1: day 3, $102.5 \geq 102.26$, upper touch; path 2: no horizontal touch in 5 days
3. Path 1: $+1$ (the later fall to 99.8 does not matter); path 2: $\ln(99.1/100) = -0.90\%$ at the vertical barrier: -1

⊡ Interpretation: the label records the first exit a trader would have taken, not the end-of-window price; a timeout keeps the sign of the terminal return, so "vertical" and "-1" are different facts

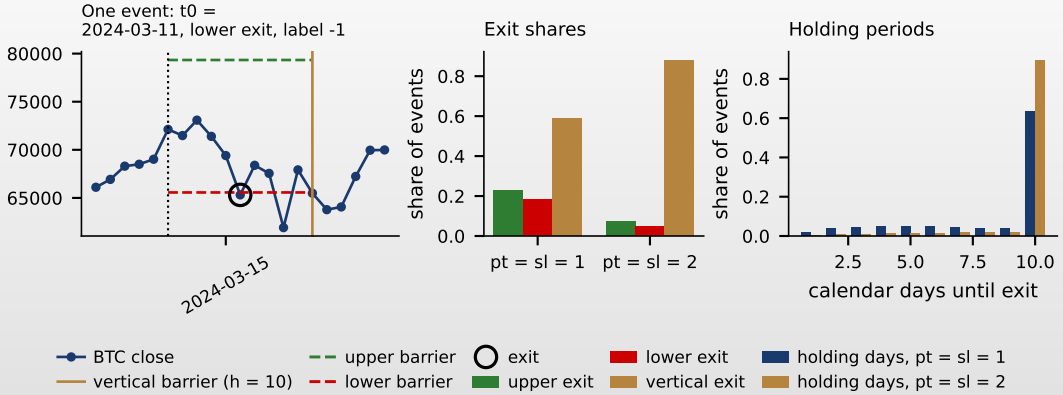


B1 [Solved]: Triple-Barrier Labels on Bitcoin — Task

- **Questions**
  - how often does Bitcoin hit a horizontal barrier within 10 days?
  - how balanced are the labels?
- **Data and settings**: BTC daily closes; every day is an event t_0; σ_{t_0} = EWMA volatility ("Fixed-Horizon Labels and EWMA Volatility"; `get_daily_vol`); $h = 10$ days; monitoring on daily closes; events without 10 later closes are dropped (4,373 events, 19 Sep 2014 – 8 Sep 2026)
- **Tasks**
  1. With `triple_barrier`, label every event for $pt = sl = 1$ and $pt = sl = 2$; print one event record (t_0, σ_{t_0}, barrier prices, t_1, exit type, return, label).
  2. Compute the shares of upper, lower and vertical exits, the average holding period and the share of $+1$ labels.
  3. Cross-tabulate exit type against the final label, and compare with the 10-day fixed-horizon label on the same events.
- **Report**: the event record, a table by setting, the cross-tabulation, the exit-share and holding-period chart
- **Interpretation**
  - why are about 56% of the labels $+1$?
  - what is then the accuracy to beat?
- Notebook: section B1

Back Exercise list

B1 [Solved]: Solution (1/2) — Triple-Barrier Labels on Bitcoin



B1 [Solved]: Solution (2/2) — Triple-Barrier Labels on Bitcoin

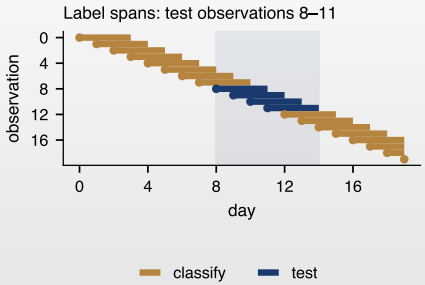
	Upper	Lower	Vertical	Days	$+1$
$pt = sl = 1$	22.8%	18.2%	59.0%	8.25	55.7%
$pt = sl = 2$	7.2%	4.9%	87.9%	9.58	55.6%
fixed 10 days	–	–	100%	10	55.5%

- Event 11 Mar 2024: close 72,124, $\sigma = 3.0\%$, barriers 79,336 and 65,568; lower exit on 16 Mar, return -9.9%, label -1
- $pt = sl = 1$: vertical exits split into 1,442 labels $+1$ and 1,139 labels -1; horizontal exits give 995 $+1$ and 797 -1
- Interpretation: BTC rose over the sample, so "always up" (about 56%) is the benchmark; wider barriers push labels towards fixed-horizon labels

MFM_ch13_seminar

A3 [Solved]: How Much Does Purging Cost? — Task

- **Question**: which training observations must go when the test labels overlap them?
- **Inputs**: $n = 20$ observations $i = 0, \ldots, 19$ (zero-based) in $K = 5$ contiguous folds of $L = 4$; the label of i spans the closed interval $[i, i + h]$, $h = 3$; embargo $e = 1$ observation after the end of the last test label; test fold $\{8, \ldots, 11\}$, whose labels reach day 14
- **Tasks**
  1. List the observations before the test fold whose span reaches day 8 (purged before).
  2. List those after the fold that start on or before day 14 (purged after) and the embargoed observation, the first one after day 14.
  3. Count the union of removals and the share of the 16 non-test observations lost; for an interior fold with $n - L$ large enough on both sides, show the share is $\frac{2h+e}{n-L}$.
- **Report**: the removed sets, the share lost, the formula
- **Interpretation**: why does shuffled K-fold overstate accuracy in this setting? Notebook: section A3



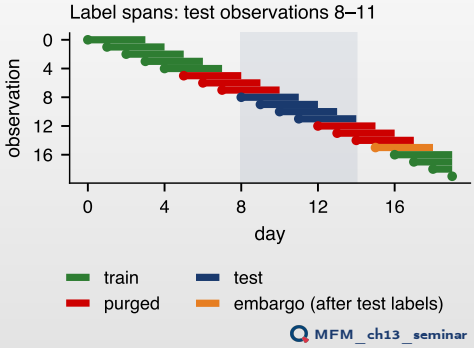
MFM_ch13_seminar

Back Exercise list

A3 [Solved]: Solution — How Much Does Purging Cost?

1. Purged before: $\{5, 6, 7\}$, since $i + 3 \geq 8$
2. Purged after: $\{12, 13, 14\}$, since they start by day 14; embargo: $\{15\}$, the first observation after day 14
3. Union: $\{5, 6, 7, 12, 13, 14, 15\}$, 7 of 16 lost (43.75%); training set $\{0, \ldots, 4\} \cup \{16, \ldots, 19\}$

- General interior fold: h observations before; after, h labels starting by the end of the last test label, then e embargoed; share $\frac{2h+e}{n-L} = \frac{3+3+1}{16}$
- Interpretation: shuffling places neighbours that share returns on both sides of the split; a flexible model memorises them (B2)



A7 [Proposed]: Purging with Shorter Labels — Task

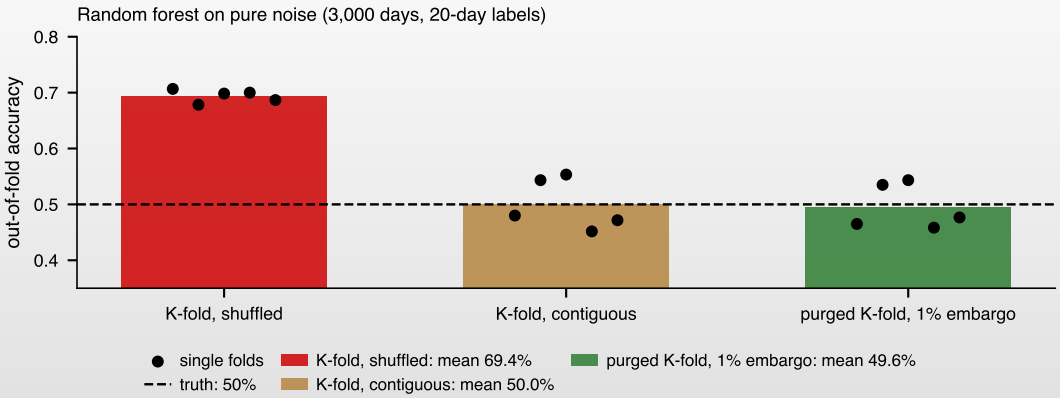
- **Model**: A3 [Solved]
- **Inputs**: as A3, with closed label spans $[i, \min(i+2, 19)]$ ($h = 2$), embargo $e = 1$, test fold $\{12, \ldots, 15\}$
- **Tasks**
  1. On a timeline of indices 0–19, fill a table with the columns purged before, embargo, purged after and union.
  2. Check the count with the A3 formula.
  3. Interior fold with $n = 250$, $K = 5$ ($L = 50$), $h = 20$, $e = 1$: compute the share lost; repeat with $K = 10$ ($L = 25$).
- **Report**: the table on the 0–19 timeline, the formula check and the shares lost for $K = 5$ and $K = 10$
- **Interpretation**
  - what does purging cost when labels are long and folds are short?
  - how would you respond?
- Notebook: section A7

B2 [Solved]: Purged K-Fold vs. Shuffled K-Fold — Task

- **Question**: how much accuracy does shuffling invent when there is nothing to predict? A synthetic diagnostic, no market data
- **Inputs** (notebook section B2, seed 42): a random walk of 3,000 days; label = sign of the 20-day forward return, spans $[i, i+20]$; 5 features = 20-day means of independent noise; true accuracy 50%
- **Tasks**
  1. Audit the supplied `PurgedKFold`: find the line that purges before the test fold, the line that applies the embargo and the one that purges after it; run it on the A3 example.
  2. Fit a random forest (Breiman (2001); 200 trees, minimum leaf 5, seed 42) under shuffled 5-fold, contiguous 5-fold and purged 5-fold with a 1% embargo (30 observations).
  3. Report the five fold accuracies and their mean for each scheme.
- **Report**: the A3 check, a chart and a table of fold accuracies
- **Interpretation**
  - in this design, what removes the leak: purging, or not shuffling?
  - would the gap shrink with 1-day labels?

Back Exercise list

B2 [Solved]: Solution (1/2) — Purged K-Fold vs. Shuffled K-Fold



B2 [Solved]: Solution (2/2) — Purged K-Fold vs. Shuffled K-Fold

- The class lines: `before = idx[t1 < test_t0]` purges before; `later = idx[(idx > stop) & (t0 > test_t1)]` finds the first observation starting after the last test label; `after_start = min(later[0] + embargo, len(X))` applies the embargo after it
- On the A3 example the class removes $\{5, 6, 7, 12, 13, 14, 15\}$ for the test fold $\{8, \ldots, 11\}$, as in the step-by-step count

Scheme	Fold accuracies; mean
shuffled	0.707 0.678 0.698 0.700 0.687; **0.694**
contiguous	0.480 0.543 0.553 0.452 0.472; 0.500
purged	0.465 0.535 0.543 0.458 0.477; **0.496**

- Purged training sets keep 2,330–2,380 of 2,400 observations
- Interpretation: here contiguous folds alone remove most of the leak
  - only the fold boundaries share returns; purging makes the boundaries clean
  - the leak needs overlapping labels *and* persistent features (Bergmeir, Hyndman & Koo (2018))
  - this one seed shows it for this design only
- With 1-day labels (same simulation) the mean accuracies are: shuffled 0.496, contiguous 0.489, purged 0.496
  - labels no longer overlap, so shuffling has nothing to memorise and the gap disappears (Bergmeir, Hyndman & Koo (2018))

MFM_ch13_seminar

The Bitcoin Modelling Frame (B3, B4, B9, B10, B14)

- Target: $y_t = 1\{\ln P_{t+\mathbf{5}} - \ln P_t > 0\}$, 5 calendar days (crypto trades daily); B3–B4 use this fixed-horizon label, whereas B1 uses triple-barrier exits; label span $[t, t+5]$
- Rows: every day with all features and a complete 5-day window: $n = 4{,}181$, 4 Apr 2015 – 13 Sep 2026, 54.8% "up"
- Validation: `PurgedKFold(5, t1, pct_embargo=0.01)`: 5 contiguous folds of 836–837 days, embargo 41 days; the same folds for every model

11 features (`build_features`)	Definition, information up to day t
`mom_1, mom_5, mom_20, mom_60, mom_120`	$x_t - x_{t-h}$
`vol_20, vol_60, vol_ratio`	standard deviation (SD) of r_t over 20 and 60 days $\times\sqrt{365}$; their ratio
`dist_ma50, dist_ma200`	x_t minus the mean of x over the last 50 or 200 days
`rsi_14`	relative strength index, 14 days, divided by 100

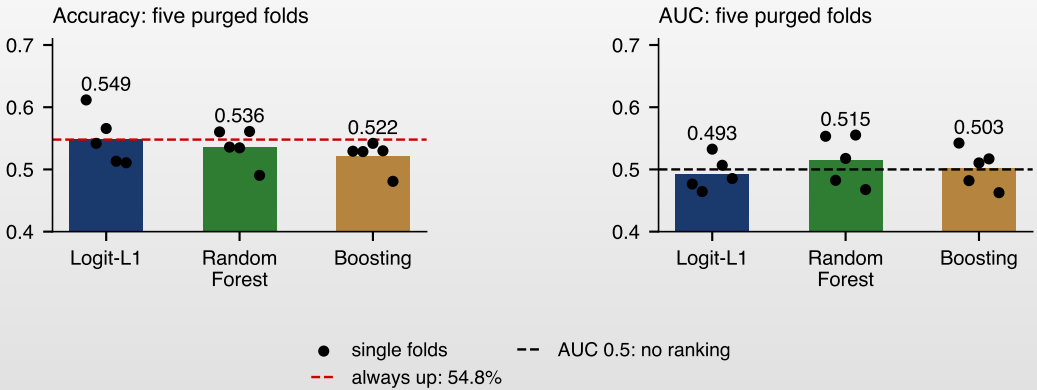
- Models: Logit-L1 (standardised, $C = 0.05$, Tibshirani (1996)); RF (400 trees, minimum leaf 50, Breiman (2001)); GB (histogram, depth 3, learning rate 0.03, 200 trees, Friedman (2001)); seed 42
- Saved for later exercises: one row per date with the label, the fold, the three out-of-fold probabilities and the constant forecast (training share of "up" in that fold)

B3a [Solved]: Three Models against "Always Up" — Task

- **Question**: does any of three models predict the 5-day direction of Bitcoin better than "always up"?
- **Data**: the Bitcoin modelling frame (previous slide)
- **Tasks**
  1. Fit Logit-L1, RF and GB on each purged training fold and predict the test fold; save the out-of-fold probabilities.
  2. For each model, compute the accuracy ($\hat{p} > 0.5$) and the ROC AUC in each test fold, and their means (**fold means**).
  3. Compute the "always up" accuracy of each test fold, its share of up labels, and the mean over folds.
  4. Compute the **pooled** AUC: one AUC over all 4,181 out-of-fold probabilities together.
- **Report**: a table (accuracy, AUC, fold SD), the baseline, the chart with fold points, the share of "up" predictions
- **Interpretation**: why is 54.8% accuracy not evidence of skill here? Notebook: section B3

Back Exercise list

B3a [Solved]: Solution (1/2) — Three Models against "Always Up"



B3a [Solved]: Solution (2/2) — Three Models against "Always Up"

	Logit-L1	RF	GB
Accuracy, fold mean	54.9%	53.6%	52.2%
AUC, fold mean	0.493	0.515	0.503
AUC, pooled	0.468	0.512	0.497
Predicted "up"	99.6%	69.5%	68.4%

- "Always up": fold shares 61.3%, 53.5%, 56.6%, 51.3%, 51.3%, mean 54.8%; RF and GB stay below it; Logit-L1 (54.9%) is in effect "always up"
- Fold mean and pooled AUC differ: five rankings averaged vs. one ranking over all days, which also mixes the folds' calibration
- Interpretation: 54.8% is the share of "up" 5-day windows, not a skill; the next question is how uncertain the RF's 0.512 is (B3b)

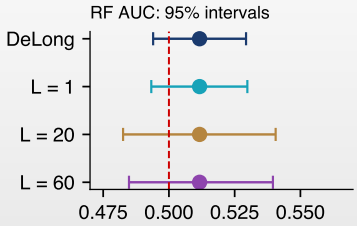
MFM_ch13_seminar

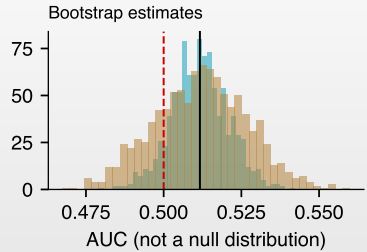
B3b [Solved]: A Dependence-Robust CI for the AUC — Task

- **Question**: is the RF's pooled AUC of 0.512 distinguishable from 0.5?
- **Data**: the saved RF out-of-fold probabilities and labels of B3a ($n = 4{,}181$); AUC as in "Accuracy, the ROC Curve and the AUC", 0.5 = no ranking
- **Tasks**
  1. Compute the DeLong standard error (DeLong, DeLong & Clarke-Pearson (1988)), the 95% CI, $z = (\mathrm{AUC} - 0.5)/\mathrm{SE}$ and the two-sided p-value.
  2. Resample the rows with a circular block bootstrap (Künsch (1989)): blocks of $L \in \{1, 20, 60\}$ consecutive days, 1,000 replications, seed 42; recompute the AUC each time.
  3. Report the bootstrap SE, the percentile 95% CI and the share of replications at or below 0.5 for each L.
- **Report**: an interval chart and the bootstrap distributions for $L = 1$ and $L = 20$
- **Interpretation**
  - why is the block-bootstrap interval wider?
  - which block length would you defend?
- Notebook: section B3

Back Exercise list

B3b [Solved]: Solution — A Dependence-Robust CI for the AUC





AUC 0.5 bootstrap L = 20
bootstrap L = 1 observed AUC

- DeLong: SE 0.0090, CI [0.494; 0.529], $z = 1.29$, $p = 0.20$; $L = 1$: SE 0.0092, CI [0.493; 0.530], as DeLong
- $L = 20$: SE 0.0150, CI [0.483; 0.541], 21.9% of replications ≤ 0.5; $L = 60$: SE 0.0141, CI [0.485; 0.540]
- Interpretation: overlapping 5-day labels and volatility clusters make neighbouring days dependent; the dependence-robust SE is 1.7 times the DeLong SE. A block of at least the label horizon, stable from 20 to 60 days: $L = 20$ is defensible. The histograms are estimator distributions, not null distributions, with the fitted models held fixed

MFM_ch13_seminar

B4 [Proposed]: Diebold–Mariano on Probability Forecasts — Task

- **Model**: B3a and B3b [Solved]
- **Question**: does any model forecast the probability of an up week better than a constant?
- **Inputs**: the saved probabilities of B3a, clipped to $[10^{-6}, 1-10^{-6}]$; log loss ℓ_t in "The Log Loss"; worked row, 13 Jul 2015, $y_t = 0$: constant $\hat{p} = 0.529$, $\ell = -\ln(1-0.529) = 0.753$; RF $\hat{p} = 0.4953$, $\ell = 0.684$
- **Tasks**
  1. Compute the mean log loss of Logit-L1, RF, GB and the constant.
  2. For RF vs. Logit-L1, RF vs. constant and GB vs. constant, form $d_t = \ell_t^A - \ell_t^B$ ($d_t > 0$: model A worse) and regress d_t on a constant with Newey–West errors (Newey & West (1987)), lags 4 and 20: the DM statistic is the t-value (Diebold & Mariano (1995)), two-sided.
  3. Compute the naive i.i.d. t of $\bar{d}$ and compare.
- **Report**: a table of log losses, $\bar{d}$, DM (lags 4, 20) with p-values, the naive t; a chart of cumulative d_t
- **Interpretation**: can a model with AUC above 0.5 lose to a constant in log loss? Pooled CV probabilities are descriptive: a deployment claim needs walk-forward forecasts. Notebook: section B4

Back Exercise list

C1 [Proposed]: Critique an AI Answer — the Answer

- **Prompt** to an AI assistant: "Predict tomorrow's direction of Bitcoin from past returns and evaluate the model out of sample with 5-fold cross-validation"
- **Answer** (`btc`: daily closes 2014-09-17 to 2026-09-18; imports omitted):

```
r = np.log(btc).diff()
X = pd.DataFrame({'r1': r, 'r5': r.rolling(5).sum(), 'vol20': r.rolling(20).std(),
                  'trend10': r.rolling(10, center=True).mean()})
data = pd.concat([X, r.shift(-1).rename('fwd')], axis=1).dropna()
y = (data['fwd'] > 0).astype(int)                # 1 if tomorrow is up
Xs = StandardScaler().fit_transform(data[X.columns])
cv = KFold(n_splits=5, shuffle=True, random_state=0)
model = LogisticRegression(penalty='l1', C=0.1, solver='liblinear')
auc = cross_val_score(model, Xs, y, cv=cv, scoring='roc_auc')
print(round(auc.mean(), 3))
```

- "Output: 0.697. An out-of-sample AUC of 0.70 on 4,361 days is a strong, tradable signal; cross-validation is out of sample, so there is no overfitting"

Back Exercise list

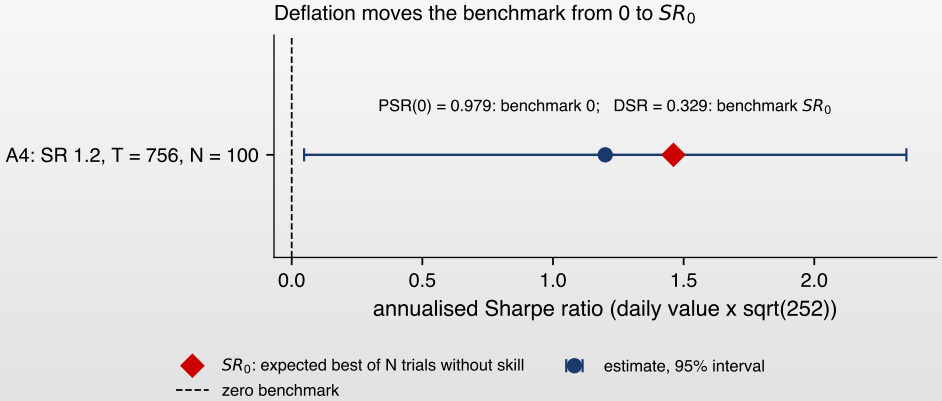
C1 [Proposed]: Critique an AI Answer — Task

- **Model**: B2 and B3a [Solved]; B9 (extension) is the same trap in a larger model
- **Tasks**
  1. Give five annotated findings, each with a line number: two leaks of test information, one validation design that does not match deployment, two unsupported claims.
  2. For each, say how you would detect it: a check on the output, a result of B2 or B3a, a rule from the section What You Need for Today.
  3. Write one corrected pipeline: trailing features, the scaler inside each training fold (`make_pipeline`), `TimeSeriesSplit(n_splits=5, gap=20)`, mean AUC over the 5 test folds.
  4. Report a before/after table: the AI code, then each correction added one at a time.
  5. Log the errors in the format of `AI_ERRORS.md`: request, answer, error, how found, correction.
- **Interpretation**
  - which days enter `trend10` on day t? Check the alignment on the toy series $0, 1, \ldots, 30$ using the same rolling call.
  - which of them is the target?
- An unusually high AUC is a reason to inspect the code, not a proof of leakage

A4 [Solved]: PSR and DSR, a Worksheet — Task

- **Question**: is an annualised Sharpe ratio of 1.2 on three years convincing if it is the best of 100 trials?
- **Inputs** (equity strategy, $A = 252$ trading days a year): annualised $\widehat{SR}_{ann} = 1.2$, $T = 756$ daily returns, $\hat{\gamma}_3 = -0.4$, ordinary kurtosis $\hat{\gamma}_4 = 5$; $N = 100$ independent trials, $V = \mathrm{Var}[\widehat{SR}_n] = 1/T$ (Bailey & López de Prado (2012); Bailey & López de Prado (2014))
- **Tasks**
  1. Compute the daily $\widehat{SR} = 1.2/\sqrt{252}$ and the denominator D in "The Probabilistic Sharpe Ratio".
  2. Compute $z = \widehat{SR}\sqrt{T-1}/D$ and $PSR(0) = \Phi(z)$.
  3. Compute the expected maximum SR_0 in "Selection and the Deflated Sharpe Ratio", daily and annualised.
  4. Compute $DSR = PSR(SR_0) = \Phi\big((\widehat{SR} - SR_0)\sqrt{T-1}/D\big)$.
- **Report**: every intermediate number, PSR(0) and DSR; the derivation of D is in the extension (A4 Extended)
- **Interpretation**
  - under i.i.d. returns, a zero benchmark and 100 trials, what does the DSR say?
  - which N would change it?
- Notebook: section A4

A4 [Solved]: Solution (1/2) — PSR and DSR, a Worksheet



A4 [Solved]: Solution (2/2) — PSR and DSR, a Worksheet

1. $\widehat{SR} = 1.2/\sqrt{252} = 0.07559$; $D = \sqrt{1 + 0.4 \cdot 0.07559 + 1 \cdot 0.07559^2} = 1.01782$
2. $z = 0.07559 \cdot \sqrt{755}/1.01782 = 2.041$; $PSR(0) = \Phi(2.041) = 0.979$
3. $\sqrt{V} = 1/\sqrt{756} = 0.03637$; $\Phi^{-1}(0.99) = 2.326$, $\Phi^{-1}(1 - \frac{1}{100e}) = 2.680$;
   $SR_0 = 0.03637\,[0.4228 \cdot 2.326 + 0.5772 \cdot 2.680] = 0.0920$ a day, 1.46 annualised
4. $z = (0.07559 - 0.0920)\sqrt{755}/1.01782 = -0.444$; $DSR = \Phi(-0.444) = 0.329$

⊡ Interpretation: luck alone gives a best annualised SR of 1.46 among 100 i.i.d. zero-skill trials

- above the observed 1.2: the evidence does not clear the illustrative 0.95 threshold
- with $N = 1$, set $SR_0 = 0$ directly (do not evaluate the large-N quantile approximation): DSR equals $PSR(0) = 0.979$

MFM_ch13_seminar

Team Project: Choose One Route

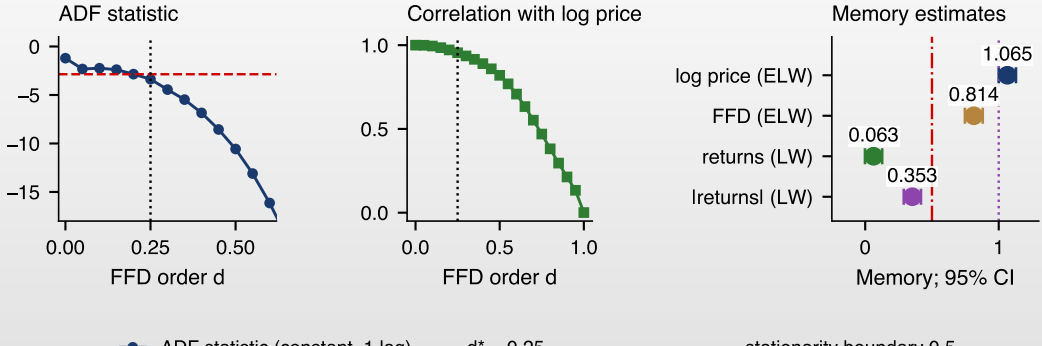
- **Standard**: one asset of your choice through B1, B3a, B3b, B5 and B6, with the candidate-return matrix of every rule tried saved
- **Advanced**: C2 (foundation model vs. boosting) or C3 (a no-arbitrage stochastic discount factor, SDF, on deciles), in the extension packs
- Minimum evidence: code as a Quantlet, one chart, the test against a benchmark, the number of trials; AI use in `AI_USE.md`, caught errors in `AI_ERRORS.md`
- State the data source and download date; freeze the sample, report its size and first/last dated close. A rigorously documented negative result is valid.
- **Assessment criteria**: labels and validation without leakage; the uncertainty of every result (CI, DSR); the number of trials reported; reproducible code; the main limitation stated

B5 [Solved]: Memory of the Bitcoin Log Price — Task

- **Questions**
  - which d does the minimum-ADF rule choose for Bitcoin?
  - what is the memory δ of the log price?
- **Data**: $x_t = \ln P_t$, BTC 2014-09-17 to 2026-09-18 ($n = 4{,}385$); supplied: `frac_diff_ffd`, `exact_local_whittle`, `local_whittle`
- **Tasks**
  1. For $d \in \{0, 0.05, \ldots, 1\}$, filter x_t ($\tau = 10^{-4}$), run `adfuller(x, maxlag=1, regression='c', autolag=None)` (Dickey & Fuller (1979)) and record the statistic and $\mathrm{Corr}(\tilde{x}_t, x_t)$ on the dates the filter keeps.
  2. Report d^*, the smallest d with a statistic below the 5% critical value.
  3. Estimate δ with 95% CIs ($m = \lfloor n^{0.65} \rfloor$ Fourier frequencies of each series, as in "Memory, the ADF Test and Local Whittle"): exact local Whittle (ELW, Shimotsu & Phillips (2005)) for x_t and the FFD series at d^*, local Whittle (LW, Robinson (1995)) for r_t and $|r_t|$.
  4. Test $H_0 : \delta = 1$ for x_t with $z = (\hat{\delta} - 1)/\mathrm{SE}$, from the unrounded estimate.
- **Report**: the ADF–correlation chart, d^*, a table of the four estimates with CIs, z and p
- **Interpretation**
  - is the FFD series at d^* stationary?
  - where does the memory of Bitcoin sit?
- Notebook: section B5

Back Exercise list

B5 [Solved]: Solution (1/2) — Memory of the Bitcoin Log Price



B5 [Solved]: Solution (2/2) — Memory of the Bitcoin Log Price

- ADF: -1.21 ($d = 0$), -2.32 (0.05), -2.24 (0.10), -2.39 (0.15), -2.85 (0.20), -3.38 (0.25); 5% value -2.86: $d^* = 0.25$, correlation 0.955; the path is not monotone in d
- ELW log price 1.065 [1.000; 1.129] ($m = 232$); FFD at d^* 0.814 [0.747; 0.880]; LW returns 0.063 [-0.002; 0.127]; $|r_t|$ 0.353 [0.289; 0.417]
- $z = (1.0647 - 1)/0.0328 = 1.97$, two-sided $p = 0.049$: a borderline rejection towards $\delta > 1$ (memory above a unit root), not towards stationarity
- Interpretation: the FFD series has $\hat{\delta} \approx 1 - d$ over the band and its CI lies far above $\frac{1}{2}$: not stationary; Bitcoin's long memory is in volatility $|r_t|$

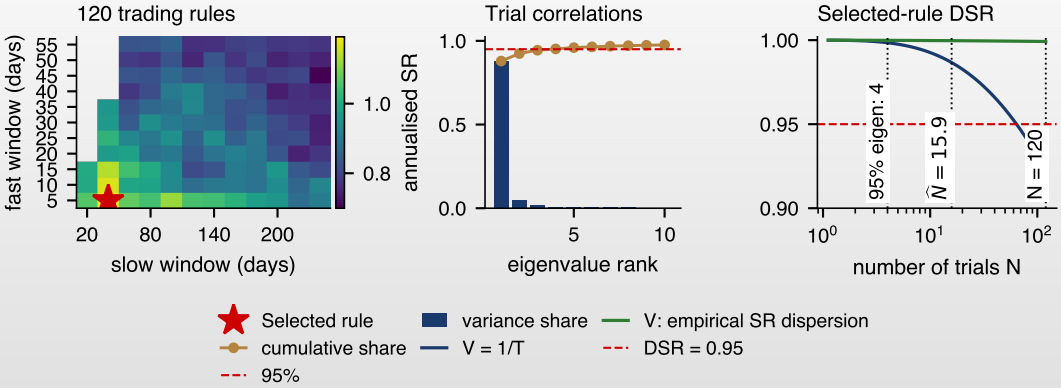
MFM_ch13_seminar

B6 [Solved]: Grid Search, DSR and the Effective Number of Trials — Task

- **Question**: after searching 120 moving-average rules on Bitcoin, is the best one better than luck?
- **Rules**: fast window $f \in \{5, 10, \ldots, 55\}$, slow $s \in \{20, 40, \ldots, 240\}$, $s > f$: $N = 120$; signal on day t: $1\{\bar{x}_{f,t} > \bar{x}_{s,t}\}$ ($\bar{x}_{f,t}$: mean of the log price over the last f days), held on day $t+1$: $r_{t+\mathbf{1}}^{(f,s)} = \text{signal}_t \cdot r_{t+\mathbf{1}}$; long/cash, gross of costs, risk-free rate 0
- **Dates**: common start on the first day every rule has a signal, 15 May 2015, to 18 Sep 2026: $T = 4{,}145$ days, $A = 365$
- **Tasks**
  1. Build the $T \times 120$ return matrix; pick the rule with the best SR; report its SR, skewness, kurtosis and PSR(0) (Bailey & López de Prado (2012)).
  2. From the 120×120 correlation matrix: eigenvalues, the number reaching 95% of the variance, and $\hat{N} = \bar{\rho} + (1 - \bar{\rho})N$ with $\bar{\rho}$ the mean off-diagonal correlation (heuristics, not theorems).
  3. Compute $SR_{\mathbf{0}}$ and the DSR (Bailey & López de Prado (2014)) for $N = 120$, $\hat{N}$ and the 95% count, with $V = 1/T$.
- **Report**: a Sharpe heatmap with the selected rule, the scree plot, a DSR table
- **Interpretation**: which N and which V would you report to an investment committee? Notebook: section B6

Back Exercise list

B6 [Solved]: Solution (1/2) — Grid Search, DSR and the Effective Number of Trials



B6 [Solved]: Solution (2/2) — Grid Search, DSR and the Effective Number of Trials

$V = 1/T$	$N = 120$	$\hat{N} = 15.9$	$N = 4$
SR_0 (ann.)	0.77	0.53	0.31
DSR	0.922	0.987	0.998

- Best rule $(5, 40)$: SR 1.19 annualised (buy-and-hold 0.78); skewness 0.13, kurtosis 12.8; PSR(0) $= 1.000$
- Execution idealised at the same close; trading at the next close, the annualised SR of rule $(5, 40)$ is 1.18
- The largest eigenvalue $\lambda_1 = 105.5$ explains 87.9%; 4 eigenvalues reach 95%; $\bar{\rho} = 0.875$, $\hat{N} = 15.9$
- Interpretation: with $V = 1/T$, counting 120 correlated trials as independent over-deflates. Read the dispersion curve at raw $N = 120$; also reducing N can count the correlation twice
  - the DSR tests "above luck", not "above buy-and-hold" (B13)
  - the winner was chosen on the full sample: a claim needs a frozen rule and an untouched holdout (Chapter 13, Bitcoin grid)

MFM_ch13_seminar

Extension Packs: What Each One Needs

Pack	Exercises	Prerequisites
1. Memory	A5, B7	A1, B5
2. Labels	A2 Extended [Solved], A6, B8	A2, B1
3. Validation and leaks	A7, B9	A3, B2, the modelling frame
4. Predictive inference	A8 [Solved], A9, B3c [Solved], B10 [Solved]	B3a, B3b, B4
5. Strategies and selection	A4 Extended [Solved], A10, B11 [Solved], B12, B13	A4, B6
6. Inference with many controls	B14	the lecture example (VIX)
7. Projects	C2, C3	Chapter 14 (C2); the lecture case study (C3)

- Each exercise has its notebook section with the same label; do them in order within a pack
- C2 and C3 are possible starts for the team project

A5 [Proposed]: Window Length and ADF Over-Rejection — Task

- **Model**: A1 [Solved]
- **Question**: how often does ADF reject a unit root in the FFD series of a pure random walk?
- **Simulation** (`rng = default_rng(42)`): for $d \in \{0.3, 0.5\}$ and each of 500 paths, compute K, draw $T + K$ innovations $\varepsilon_s \sim N(0,1)$ with $T = 2{,}500$, set $X_t = \sum_{s \le t} \varepsilon_s$ (X starts at the first draw), apply the finite filter and keep exactly T values; then repeat with $d = 0.25$ and with no filter (control)
- **Tasks**
  1. Run `adfuller(f, maxlag=1, regression='c', autolag=None)` on each path and record rejections at 5%.
  2. On the same paths, run the test with lags chosen by the Akaike Information Criterion (AIC, `autolag='AIC'`), using the statsmodels default maximum: the ceiling of $12(T/100)^{1/4}$, where T is the filtered sample length.
  3. Report each rejection frequency with its Monte Carlo SE $\sqrt{\hat{p}(1-\hat{p})/500}$.
- **Interpretation**: every path is $I(1)$
  - why does a one-lag ADF confuse a scaled random walk with a stationary series?
  - which d would make an $I(1)$ price stationary?
- Notebook: section A5

Back Exercise list

B7 [Proposed]: Robustness of the Minimum-ADF Rule for Ether — Task

- **Model**: B5 [Solved]
- **Question**: how robust is the minimum-ADF rule on an asset that grew about 1,000-fold?
- **Data**: ETH-USD log closes; three specifications, each with a full d-grid scan $\{0, 0.05, \ldots, 1\}$ and one ADF lag

Row	Sample	Deterministic terms	5% critical value
1	2015-08-07 – 2026-09-18 ($n = 4{,}061$)	constant	-2.86
2	2015-08-07 – 2026-09-18	constant + trend	-3.41
3	2018-01-01 – 2026-09-18 ($n = 3{,}183$)	constant	-2.86

- **Tasks**
  1. For each row, find d^* and record the ADF values at $d = 0$ and $d = 0.05$.
  2. Estimate ELW with a 95% CI for the log price on rows 1 and 3.
  3. Estimate LW with a 95% CI for $|r_t|$ on the full sample.
- **Report**: a table of d^* and the two ADF values for each row; the two ELW estimates and the LW estimate, with CIs
- **Interpretation**
  - which d would you use in a model?
  - why that d?
- Notebook: section B7

Back Exercise list

A2 Extended [Solved]: Hitting Probabilities — Task

- **Question**: what share of events should end at the vertical barrier if prices were a Brownian motion?
- **Setting**: log price σW_t, W_t a standard Brownian motion (no drift), barriers $+a = pt\,\sigma\sqrt{h}$ and $-b = -sl\,\sigma\sqrt{h}$; τ = first time either is hit; martingales W_t and $W_t^2 - t$ (processes whose expected future value is the current one); stopping time τ with $\mathbb{E}[\tau] < \infty$; optional stopping: then $\mathbb{E}[W_\tau] = 0$ and $\mathbb{E}[W_\tau^2] = \mathbb{E}[\tau]$ (Karatzas & Shreve (1998))
- **Tasks**
  1. Without a vertical barrier, derive P(upper first) and $\mathbb{E}[\tau]$ by optional stopping.
  2. For $pt = sl = 1$ and $h = 10$, compute $P(\tau > h)$ under continuous monitoring with $P(\tau > h) = \frac{4}{\pi}\sum_{n\geq 0}\frac{(-1)^n}{2n+1}e^{-(2n+1)^2\pi^2/8}$.
  3. Approximate daily monitoring by moving each barrier out by $0.5826\,\sigma$ (Broadie, Glasserman & Kou (1997)); then simulate 400,000 paths of 10 Gaussian daily steps (seed 42) and compare.
- **Report**: P(upper first) and $\mathbb{E}[\tau]$; $P(\tau > h)$ under continuous monitoring, approximated and simulated, with the simulation SE
- **Interpretation**: B1 finds 59.0% timeouts on Bitcoin: which part of the gap to the Brownian value is monitoring, which is the data?

Back Exercise list

A2 Extended [Solved]: Solution — Hitting Probabilities

1. W_t is a martingale: $0 = p\,a - (1-p)\,b \Rightarrow p = \frac{b}{a+b} = \frac{sl}{pt+sl}$; $\sigma^2 W_t^2 - \sigma^2 t$ is a martingale: $\sigma^2 \mathbb{E}[\tau] = p\,a^2 + (1-p)\,b^2 = ab \Rightarrow \mathbb{E}[\tau] = pt \cdot sl \cdot h$ days
2. Continuous monitoring, exact: $P(\tau > h) = 1.2732 \cdot 0.2912 - \cdots = 0.371$
3. Daily monitoring, approximate: $c = 1 + 0.5826/\sqrt{10} = 1.184$, replacing $\pi^2/8$ by $\pi^2/(8c^2)$ gives 0.528; simulated with daily steps: 0.517 (SE 0.001)

⊡ Interpretation: discrete monitoring explains most of the gap from 37% to 59%; the remaining 7 points may come from fat tails, volatility clustering and error in the EWMA volatility, which this comparison does not separate

A6 [Proposed]: Asymmetric Barriers — Task

- **Model**: A2 and A2 Extended [Solved]
- **Inputs**: $P_{t_0} = 50$, $\sigma_{t_0} = 2\%$, $h = 4$, $pt = 2$, $sl = 1$; paths (i) 50, 51.5, 53.0, 49.5, 47.9 and (ii) 50, 52.0, 54.5, 53.0, 51.0
- **Tasks**
  1. Compute both barriers, draw them on the two paths and label each path. A horizontal breach on the last day takes precedence over the vertical exit.
  2. Optional (drift): for $X_t = \mu t + \sigma W_t$ (μ: drift per day) without a vertical barrier, the scale function $s(x) = e^{-2\mu x/\sigma^2}$ solves $\frac{\sigma^2}{2}s'' + \mu s' = 0$, so $s(X_t)$ is a martingale; derive $P(\text{upper first}) = \frac{e^{2\mu b/\sigma^2} - 1}{e^{2\mu b/\sigma^2} - e^{-2\mu a/\sigma^2}}$ and evaluate it with $pt = 2$, $sl = 1$, now changing the daily volatility to $\sigma = 3\%$, $\mu = 0.1\%$ a day, $h = 10$.
- **Report**: the barriers, the labels, the formula and its value
- **Interpretation**: the formula has no vertical barrier: which paths would a vertical barrier at h censor? Notebook: section A6

Back Exercise list

B8 [Proposed]: Asymmetric Barriers on Bitcoin — Task

- **Model**: B1 [Solved] and A2 Extended
- **Data**: the events, EWMA volatility and labelling of B1, $h = 10$; settings $(pt, sl) = (2, 1)$ and $(1, 2)$
- **Tasks**
  1. Report the exit shares, the share of $+1$ labels and the average holding period for both settings.
  2. Compute the upper share among horizontal exits, upper/(upper + lower).
  3. Compare it with $sl/(pt + sl)$ (no vertical barrier) and with a simulation: 400,000 paths of 10 Gaussian daily steps in units of σ, barriers at $pt\sqrt{10}$ and $-sl\sqrt{10}$, seed 42, zero drift and the BTC drift $\bar{r}/\hat{\sigma} = 0.034$ a day; give the simulation SE.
- **Report**: a table of exit shares, the $+1$ share and the mean holding period for both settings; the up share among horizontal exits, with its theoretical and simulated values, with SEs
- **Interpretation**: how does the choice change the class balance a classifier sees? The asymmetry is a modelling hypothesis, not an optimal trading rule. Notebook: section B8

Back Exercise list

B9 [Proposed]: Find the Leak — Task

- **Model**: B2 [Solved] and C1
- **Data**: the modelling frame, plus a 12th feature built in the notebook as `ma11 = r.rolling(11, center=True).mean()`; RF with 300 trees (not 400 as in B3a), minimum leaf 50, seed 42, purged 5-fold; 4,181 rows
- **Tasks**
  1. Write the dates $t-5, \ldots, t+5$ that enter `ma11` on day t and mark those inside the target window $t+1, \ldots, t+5$; rank the 12 features by MDI from one random forest fitted on all rows (a descriptive diagnostic; prediction scores remain out-of-fold).
  2. Compute the correlation of `ma11` and of the trailing `r.rolling(11).mean()` with the 5-day forward return.
  3. Report accuracy and AUC for: 11 features + centred `ma11`; 11 + trailing; the 11 base features, same rows, folds and settings.
- **Interpretation**: why can purging not catch this leak? Give a second example of look-ahead in a feature pipeline.

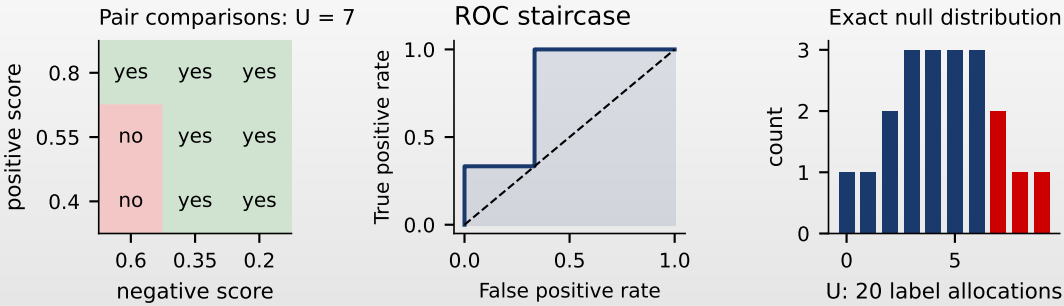
Notebook: section B9

A8 [Solved]: AUC Step by Step — Task

- **Questions**
  - what does an AUC measure?
  - how uncertain is it with six observations?
- **Inputs**: scores of 3 positives (0.80, 0.55, 0.40) and 3 negatives (0.60, 0.35, 0.20); $n_1 = n_0 = 3$ positives and negatives; U = number of (positive, negative) pairs ranked correctly, $A = U/(n_1 n_0) = U/9$; Hanley–McNeil SE (Hanley & McNeil (1982)): $\mathrm{SE}^2 = \frac{A(1-A)+(n_1-1)(Q_1-A^2)+(n_0-1)(Q_2-A^2)}{n_1 n_0}$, $Q_1 = \frac{A}{2-A}$, $Q_2 = \frac{2A^2}{1+A}$
- **Tasks**
  1. Fill the 3×3 comparison grid (is the positive above the negative?) and compute U and A.
  2. Draw the ROC staircase and check that its area equals A.
  3. Compute the SE and the Wald $z = (A - 0.5)/\mathrm{SE}$, two-sided.
  4. Enumerate the 20 ways to label 3 of the 6 scores as positive, compute U for each and the exact $P(U \geq 7)$.
- **Report**: the grid, U, A, the ROC curve, the SE, z, p and the exact $P(U \geq 7)$
- **Interpretation**: which assumption of the SE fails for overlapping 5-day labels? Notebook: section A8

Back Exercise list

A8 [Solved]: Solution (1/2) — AUC Step by Step



A8 [Solved]: Solution (2/2) — AUC Step by Step

1. 0.80 beats 3 negatives, 0.55 beats 2, 0.40 beats 2: $U = 7$, $A = 0.778$; (2) the staircase has area $\frac{1}{3} \cdot \frac{1}{3} + \frac{2}{3} \cdot 1 = \frac{7}{9}$
3. $Q_1 = 0.636$, $Q_2 = 0.681$, $\mathrm{SE} = 0.207$, $z = 1.34$, $p = 0.18$ (asymptotic, unreliable with six points)
4. Exact: $P(U \geq 7) = 4/20 = 0.20$, two-sided $p = 0.40$ under exchangeable labels: six observations cannot separate skill from luck

⊡ Both SEs assume independent observations; overlapping labels share returns, so the effective sample is smaller: B3b uses a block bootstrap instead

MFM_ch13_seminar

A9 [Proposed]: AUC as a U-Statistic and the DeLong Variance — Task

- **Model**: A8 [Solved]
- **Inputs**: 4 positives (0.70, 0.62, 0.48, 0.45), 4 negatives (0.66, 0.50, 0.40, 0.30); kernel $\psi(s_i, s_j) = 1\{s_i > s_j\} + \frac{1}{2}1\{s_i = s_j\}$, with s_i the score of positive i, s_j of negative j, $1\{\cdot\}$ the indicator; $n_1 = n_0 = 4$
- **Tasks**
  1. Fill the 4×4 grid of ψ; compute U, the AUC and its Hanley–McNeil SE.
  2. Row means $V_{10}(i)$ and column means $V_{01}(j)$ of the grid (placement values); their sample variances S_{10}, S_{01} with denominators $n_1 - 1$ and $n_0 - 1$.
  3. The DeLong SE $= \sqrt{S_{10}/n_1 + S_{01}/n_0}$ (DeLong, DeLong & Clarke-Pearson (1988)).
  4. Optional proof: by the Hoeffding decomposition the AUC is asymptotically the sum of two independent averages of placement values.
- **Report**: the grid, U, the AUC, the placement values, S_{10}, S_{01} and both SEs
- **Interpretation**: the argument is asymptotic, not an exact small-sample guarantee: what does that mean with 8 scores? Notebook: section A9

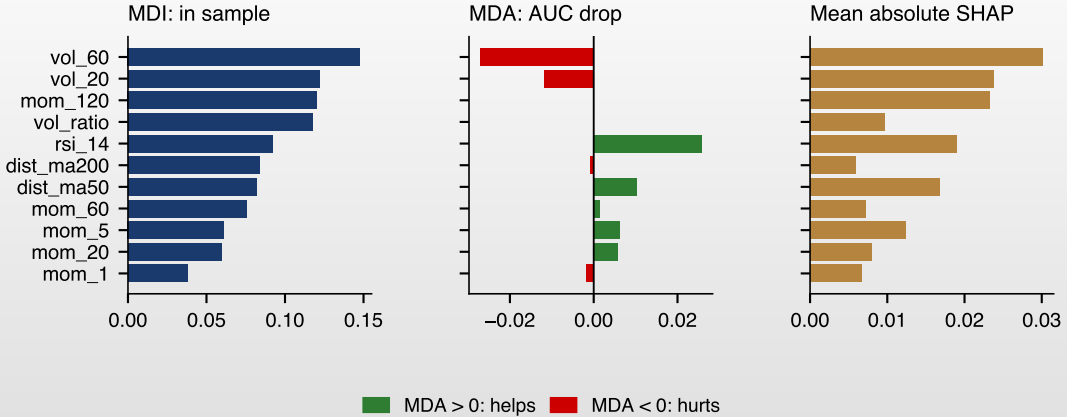
Back Exercise list

B3c [Solved]: Do Importance Measures Agree? — Task

- **Question**: do in-sample and out-of-sample importance measures rank the 11 features the same way?
- **Data**: the modelling frame, with a separate chronological split: RF (400 trees, minimum leaf 50) fitted on the first 70% (to 8 Apr 2023), evaluated after a gap of 5 + 41 days, from 24 May 2023; a single split, because importances need one fitted model
- **Tasks**
  1. MDI from the fitted random forest; MDA as the drop in test AUC when a feature is permuted (10 repeats); mean |SHAP| on 300 test days (Lundberg & Lee (2017)).
  2. Spearman rank correlations between the three measures.
- **Report**: a chart of the three measures by feature and the table of Spearman correlations
- **Interpretation**
  - what does a negative MDA mean?
  - why do correlated features blur all three measures?
- Notebook: section B3

Back Exercise list

B3c [Solved]: Solution (1/2) — Do Importance Measures Agree?



B3c [Solved]: Solution (2/2) — Do Importance Measures Agree?

- MDI ranks `vol_60`, `vol_20`, `mom_120` first; MDA is negative for both volatility features and largest for `rsi_14` (0.026)
- Spearman: MDI–SHAP 0.75, MDI–MDA -0.42, MDA–SHAP -0.10: in-sample and out-of-sample rankings disagree
- Interpretation: a negative MDA means the permuted feature scores better than the real one on new data: a sign of overfitting, within sampling noise; correlated features share and dilute importance, and permutation extrapolates off the data

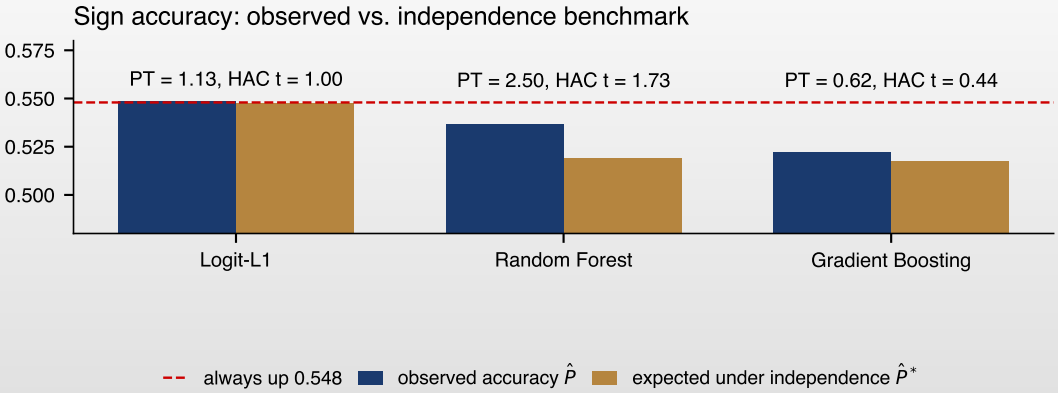
MFM_ch13_seminar

B10 [Solved]: Pesaran–Timmermann and Clark–West — Task

- **Questions**
  - do the predicted signs carry information?
  - do the features add information to a constant forecast?
- **Data**: the saved probabilities of B3a and the constant of B4; $x_t = 1\{\hat{p}_t > 0.5\}$. Caveat before you start: Clark–West (CW, Clark & West (2007)) is derived for nested linear models and genuine out-of-sample forecasts; on penalised-logit probabilities from pooled purged CV it is an approximate, descriptive adjustment
- **Tasks**
  1. For each model: $\hat{P}$ = share of correct signs, $\hat{P}^* = \hat{p}_y\hat{p}_x + (1-\hat{p}_y)(1-\hat{p}_x)$ and the Pesaran–Timmermann statistic (PT), $PT = (\hat{P} - \hat{P}^*)/\sqrt{\hat{V}(\hat{P}) - \hat{V}(\hat{P}^*)}$ (Pesaran & Timmermann (1992)); $\hat{p}_y$, $\hat{p}_x$: shares of realised and predicted "up"; n: number of days; $\hat{V}(\hat{P}) = \hat{P}^*(1-\hat{P}^*)/n$, $\hat{V}(\hat{P}^*) = [(2\hat{p}_y - 1)^2\hat{p}_x(1-\hat{p}_x) + (2\hat{p}_x - 1)^2\hat{p}_y(1-\hat{p}_y)]/n + 4\hat{p}_y\hat{p}_x(1-\hat{p}_y)(1-\hat{p}_x)/n^2$.
  2. The slope t of $y_t = a + b\,x_t + u_t$ with Newey–West errors (4 lags), two-sided.
  3. Logit-L1 ($\hat{p}_{\mathbf{1},t}$) vs. constant ($\hat{p}_{\mathbf{0},t}$): $f_t = (y_t - \hat{p}_{\mathbf{0},t})^2 - [(y_t - \hat{p}_{\mathbf{1},t})^2 - (\hat{p}_{\mathbf{0},t} - \hat{p}_{\mathbf{1},t})^2]$; CW = HAC t of $\bar{f}$, one-sided (positive = the features help); DM on the Brier loss, two-sided.
- **Report**: a table of $\hat{P}$, $\hat{P}^*$, PT and t_{HAC} by model; CW and DM with their p-values
- **Interpretation**: why can DM and Clark–West disagree for nested models? Notebook: section B10

Back Exercise list

B10 [Solved]: Solution (1/2) — Pesaran–Timmermann and Clark–West



B10 [Solved]: Solution (2/2) — Pesaran–Timmermann and Clark–West

	Pred. up	$\hat{P}$	$\hat{P}^*$	PT	t_{HAC}
Logit-L1	99.6%	0.549	0.548	1.13	1.00
RF	69.5%	0.536	0.519	2.50	1.73
GB	68.4%	0.522	0.518	0.62	0.44

- RF: accuracy 53.6% is below "always up" (54.8%); $PT = 2.50$ rejects independent signs only if the days are independent; with HAC errors $t = 1.73$, $p = 0.08$: not significant at 5%
- Clark–West, Logit-L1 vs. constant: $t = -0.44$, one-sided $p = 0.67$; DM on Brier: $t = 1.27$ (positive = logit worse), $p = 0.20$
- DM asks whether the *estimated* logit forecasts better, and its estimation noise makes it lose under the nested null; CW adds back $(\hat{p}_0 - \hat{p}_1)^2$ and asks whether the population coefficients are zero; here both find no signal
- Sign dependence (PT) is not the same as beating a majority baseline, nor as a better log loss (B4)

MFM_ch13_seminar

A4 Extended [Solved]: The Variance of the Sharpe Ratio — Task

- **Question**: where does the denominator D of A4 come from?
- **Setting**: T i.i.d. returns with mean μ, variance σ^2, skewness γ_3, ordinary kurtosis γ_4; $SR = \mu/\sigma = g(m_1, m_2) = m_1/\sqrt{m_2 - m_1^2}$ with $m_j = \mathbb{E}[r^j]$; delta method: $T\,\mathrm{Var}(\widehat{SR}) \approx \nabla g^\top \Sigma \nabla g$, with ∇g the gradient of g in (m_1, m_2) and Σ the covariance matrix of (r, r^2)
- **Tasks**
  1. Compute the gradient ∇g.
  2. Compute $\mathrm{Var}(r)$, $\mathrm{Cov}(r, r^2)$ and $\mathrm{Var}(r^2)$ in terms of $\mu, \sigma, \gamma_3, \gamma_4$.
  3. Show that $T\,\mathrm{Var}(\widehat{SR}) \approx \nabla g^\top \Sigma \nabla g = 1 - \gamma_3 SR + \frac{\gamma_4 - 1}{4} SR^2$ (Lo (2002)).
- **Report**: the gradient, the three moments and the step-by-step derivation
- **Interpretation**: which moment makes a strategy with negative skewness look more precise than it is?

Back Exercise list

A4 Extended [Solved]: Solution — The Variance of the Sharpe Ratio

1. $\nabla g = \left(\frac{1}{\sigma} + \frac{\mu^2}{\sigma^3},\ -\frac{\mu}{2\sigma^3}\right)$
2. $\mathrm{Var}(r) = \sigma^2$, $\mathrm{Cov}(r, r^2) = \sigma^3\gamma_3 + 2\mu\sigma^2$, $\mathrm{Var}(r^2) = \sigma^4(\gamma_4 - 1) + 4\mu\sigma^3\gamma_3 + 4\mu^2\sigma^2$
3. With $S = \mu/\sigma$ and $(a, b) = \nabla g$, the three products are

$$a^2\mathrm{Var}(r) = 1 + 2S^2 + S^4,$$
$$2ab\mathrm{Cov}(r, r^2) = -S\gamma_3 - 2S^2 - S^3\gamma_3 - 2S^4,$$
$$b^2\mathrm{Var}(r^2) = S^2(\gamma_4 - 1)/4 + S^3\gamma_3 + S^4.$$

4. In $\nabla g^\top \Sigma \nabla g$ the terms in μ^2/σ^2 cancel together with the cubic and quartic terms and leave $1 - \gamma_3 SR + \frac{\gamma_4 - 1}{4} SR^2$; for the Normal distribution ($\gamma_3 = 0$, $\gamma_4 = 3$) $1 + \frac{1}{2} SR^2$

⊡ Interpretation: with $SR > 0$, negative skewness adds $|\gamma_3| SR$ to the variance and fat tails add $\frac{\gamma_4 - 3}{4} SR^2$: ignoring them understates the SE; A4 used $\gamma_3 = -0.4$, $\gamma_4 = 5$

A10 [Proposed]: DSR on a Shorter Sample and the Gumbel Constant — Task

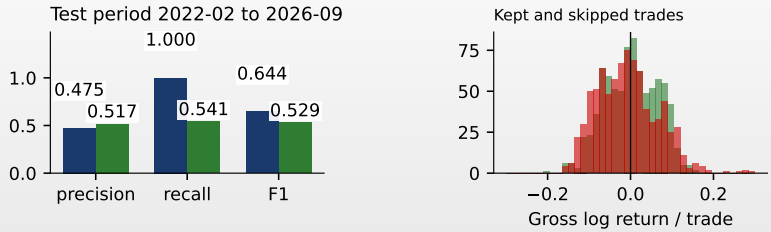
- **Model**: A4 and A4 Extended [Solved]
- **Inputs**: $\widehat{SR}_{ann} = 1.5$, $T = 504$ daily returns ($A = 252$ a year), $\hat{\gamma}_3 = -0.2$, $\hat{\gamma}_4 = 6$, $N = 20$ independent trials, $V = 1/T$; Euler's constant $\gamma_E = 0.5772$ (not the skewness)
- **Tasks**
  1. Fill the A4 worksheet: $\widehat{SR}$, D, $PSR(0)$, SR_0 (daily and annualised), DSR.
  2. Optional: with location $b_N = \Phi^{-1}(1 - \frac{1}{N})$ and scale $a_N = \Phi^{-1}(1 - \frac{1}{Ne}) - b_N$, the maximum $\max_n Z_n$ of N independent standard Normal draws Z_n is approximately Gumbel. Use $\mathbb{E}[M] = b + \gamma_E a$ for $M \sim \mathrm{Gumbel}(b, a)$; show that its mean is $(1 - \gamma_E) b_N + \gamma_E \Phi^{-1}(1 - \frac{1}{Ne})$.
  3. Optional: check it by simulation for $N = 100$, 20,000 draws, seed 42, and plot the distribution.
- **Report**: the completed worksheet, the Gumbel mean and the simulated mean with its SE
- **Interpretation**
  - is the evidence stronger than in A4?
  - does it clear 0.95?
- Notebook: section A10

B11 [Solved]: Meta-Labelling a Moving-Average Rule — Task

- **Question**: can a classifier tell when to trust a simple trend rule?
- **Primary side**: $s_t = \text{sign}(MA_{20}(P)_t - MA_{50}(P)_t)$ on prices, one candidate trade on every eligible day (a standing daily side, not only crossings); exit by the triple barrier $pt = sl = 1$, $h = 10$; gross meta-label $1\{s_t \cdot r_{t_0,t_1} > 0\}$ (López de Prado (2018)); $MA_k(P)_t$: mean close over the last k days; r_{t_0,t_1}: log return from entry to exit
- **Secondary model**: RF (400 trees, minimum leaf 50, balanced class weights) on the 11 features plus s_t; train on the first 60% of 4,174 events (to 9 Feb 2022), gap 10 days, test 20 Feb 2022 – 6 Sep 2026 (1,660 trades)
- **Tasks**
  1. Compute the confusion matrix of the rule "take the trade if $\hat{p} > 0.5$".
  2. Compute precision, recall and F1 before filtering (every trade taken) and after.
  3. Compute the gross return per trade for the kept and for the skipped trades.
- **Report**: the confusion matrix; a table of precision, recall and F1 before and after; the return per trade of the two groups
- **Interpretation**: trades overlap: this is not a portfolio backtest; what would a capital rule and costs add? Notebook: section B11

Back Exercise list

B11 [Solved]: Solution — Meta-Labelling a Moving-Average Rule



1. Taken: 426 winners, 398 losers; skipped: 362 winners, 474 losers (824 of 1,660 kept, 49.6%)
2. Before: precision 0.475, recall 1 (every winner is taken), F1 0.644; after: 0.517, 0.541, 0.529
3. Gross log return per trade: all -0.27%, kept $+0.19\%$, skipped -0.73%

⊡ Interpretation: filtering raises precision and the return per trade while F1 falls; the next steps are a block-bootstrap test of that return, costs, a capital rule for overlapping trades and a DSR over all meta-model settings tried

MFM_ch13_seminar

B12 [Proposed]: Probability of Backtest Overfitting — Task

- **Model**: B6 [Solved]; PBO (Probability of Backtest Overfitting) is defined in task 3
- **Data**: the 4,145 × 120 return matrix of B6, split into $S = 16$ contiguous blocks with `np.array_split` (one block of 260 days, fifteen of 259) (Bailey, Borwein, López de Prado & Zhu (2017))
- **Warm-up**: use the first four strategy columns and the first four contiguous blocks of the supplied return matrix; for each of the six choices of two training blocks, report the selected winner, its rank on the other two blocks, and the resulting PBO; compute Sharpe ratios from the selected daily returns
- **Tasks**
  1. For each of the $\binom{16}{8} = 12{,}870$ choices of 8 in-sample blocks, pick the rule with the best in-sample SR.
  2. Rank all 120 rules out of sample in ascending order (average ranks for ties), $\omega = \text{rank}/(N+1)$, $\lambda = \ln \frac{\omega}{1-\omega}$.
  3. Report $\text{PBO} = P(\lambda \leq 0)$, the median λ, the median in-sample (IS) and out-of-sample (OOS) SR of the winner, the share with OOS SR < 0, and the histogram of λ.
- **Interpretation**: PBO is the share of selected winners at or below the OOS median, not the probability of losing money; is the grid overfit? Notebook: section B12

Back Exercise list

B13 [Proposed]: Reality Check, SPA and StepM on the 120 Rules — Task

- **Model**: B6 [Solved]; supplied: `stationary_bootstrap_index` and `snooping_tests` (notebook section B13)
- **Null**: $H_0 : \max_k \mathbb{E}[d_{k,t}] \leq 0$ with $d_{k,t} = r_{k,t} - r_{0,t}$; benchmarks cash ($r_0 = 0$) and buy-and-hold ($r_0 = r_t$); these tests compare *mean* returns, the DSR concerns Sharpe ratios
- **Bootstrap** (Politis & Romano (1994)): one index sequence per replication, applied to every column (whole days), mean block 20 days (each day a block ends with probability 1/20), 2,000 replications, seed 42; $\hat{\omega}_k = \sqrt{n} \cdot$ bootstrap SD of $\bar{d}_k$
- **Tasks**
  1. Reality Check (RC, White (2000)), with $\bar{d}_k^*$ the mean of rule k in a replication and n the number of days: $p =$ share of replications with $\max_k \sqrt{n}(\bar{d}_k^* - \bar{d}_k) \geq \max_k \sqrt{n}\,\bar{d}_k$.
  2. SPA (Superior Predictive Ability, Hansen (2005)): studentised, $t_k = \sqrt{n}\,\bar{d}_k / \hat{\omega}_k$, recentring only rules with $t_k \geq -\sqrt{2 \ln \ln n}$; StepM (Romano & Wolf (2005)) at 5%: reject every rule above the 95% quantile of the max over the active rules, repeat until none.
  3. Report per benchmark: the largest studentised mean, both p-values, the number of nulls rejected; plot the bootstrap null of the maximum.
- **Interpretation**
  - which method handles correlated trials without choosing N?
  - does any rule beat holding Bitcoin?

Back Exercise list

B14 [Proposed]: Double Selection: Does the VIX Predict Bitcoin? — Task

- **Model**: B10 [Solved] (regression with HAC errors); the penalty and double selection are defined below
- **Caveat first**: the plug-in penalty assumes independent scores; with overlapping 5-day outcomes the interval below is an illustrative approximation, not a validated 95% CI
- **Data**: the modelling frame ($n = 4{,}181$); y_t = 5-day forward BTC log return in basis points (bp; 1 bp = 0.01%); d_t = log VIX (CBOE Volatility IndeX) close of the same calendar date, carried forward over weekends and US (United States) holidays (the last close known at the BTC close), standardised; controls: the 11 features, their squares and pairwise products ($p = 77$), standardised
- **Penalty** (Belloni, Chernozhukov & Hansen (2014)): $\min_\beta \frac{1}{n}\sum_t (y_t - x_t^\top \beta)^2 + \frac{\lambda}{n}\sum_j \hat{\psi}_j |\beta_j|$, $\lambda = 2c\sqrt{n}\,\Phi^{-1}(1 - \gamma/(2p))$, $c = 1.1$, $\gamma = 0.1/\ln n$, loadings $\hat{\psi}_j$ (the noise scale of control j) iterated; supplied as `rlasso`
- **Tasks**
  1. Fit a LASSO (Least Absolute Shrinkage and Selection Operator) of y_t on (d_t, x_t): is the VIX kept?
  2. Find $\hat{S}_y$ and $\hat{S}_d$, the controls kept by a LASSO of y_t and of d_t on x_t.
  3. Fit OLS (Ordinary Least Squares) of y_t on d_t and $x_{\hat{S}_y \cup \hat{S}_d}$ with HAC errors (4 lags): $\hat{\theta}$ (the coefficient of d_t), its SE and interval; repeat OLS with the VIX alone and with all 77 controls.
- **Report**: the selected controls; a table of $\hat{\theta}$, SE and interval for the three regressions
- **Interpretation**: why does the single LASSO give no valid inference? Read $\hat{\theta}$ as a predictive coefficient, not a causal fear premium. Notebook: section B14

Back Exercise list

C2 [Proposed]: Does a Foundation Model Beat Boosting? — Project Brief

- **Question**: does a time-series foundation model of Chapter 14, used zero-shot, forecast the 5-day direction of Bitcoin better than gradient boosting and a constant?
- **Protocol** (a team project, graded on protocol and evidence, not on the sign of the result)
  1. Name the model and checkpoint, its documented pre-training cut-off, and test only on later dates; otherwise state contamination as a limitation.
  2. Forecast the distribution of the 5-day return directly (not summed daily quantiles); $P(r_{t,t+\mathbf{5}} > 0)$ by linear interpolation of the predictive distribution function at 0, stated tail treatment.
  3. Refit gradient boosting and the constant at every origin on a fixed rolling window (e.g. 3 years); score log losses on identical dates; Giacomini–White test (Giacomini & White (2006)) with HAC errors; a comparison chart.
  4. Optional: a long/cash rule on $P(\text{up}) > 0.5$ with costs, and its DSR over all configurations and prompts tried.
- **Deliverables**: a Quantlet folder (code, `Metainfo.txt`, one chart) and one slide with the conclusion and its main limitation

C3 [Proposed]: A No-Arbitrage SDF on Deciles — Project Brief (1/2)

- **Question**: does a linear no-arbitrage stochastic discount factor (SDF) on characteristic-sorted portfolios beat the Fama–French tangency portfolios (Fama & French (2015)) out of sample, as in the lecture case study (Chen, Pelger & Zhu (2024))?
- **Data** (archive and column manifest): value-weighted decile portfolios of 15 univariate sorts of the Kenneth French Data Library (size, book-to-market, operating profitability, investment, earnings/price, cash flow/price, dividend yield, momentum, short- and long-term reversal, accruals, market beta, net share issues, variance, residual variance), monthly, minus the one-month bill rate; the FF3 (Fama–French three-factor model) and FF5 (Fama–French five-factor model) factor files
- **Legs**: per characteristic, the mean of deciles 1–5 and of deciles 6–10: 30 factors $\tilde{F}_{t+1}$; months with all 30 legs and FF5 only, no imputation; save the frozen panel with its download date
- **Caution**: the pricing conditions and an unconstrained linear SDF do not by themselves rule out arbitrage; positivity of m_{t+1} is not imposed
- **Limitation**: halves of value-weighted deciles replace the rank-weighted legs of 46 characteristics on individual stocks: an extension on portfolio data, not a replication of the stock-level GAN (generative adversarial network) model
- **Split**: estimate on 1967–1986, tune on 1987–1991, test on 1992–2016 and on 2017 to the latest month

Back Exercise list

C3 [Proposed]: A No-Arbitrage SDF on Deciles — Project Brief (2/2)

- **Model**: A4 (Sharpe ratio inference) and B10 (predictive tests) [Solved]
- **Stages**
  1. Reproduce the FF3 and FF5 benchmarks: tangency weights $w \propto \hat{\Sigma}^{-1}\hat{\mu}$ on 1967–1986, monthly SR on 1992–2016 with a 95% CI (Lo (2002)).
  2. With $\mu =$ mean of $\tilde{F}$ and $Q =$ mean of $\tilde{F}\tilde{F}^\top$ (a second moment, not a covariance) on 1967–1986: $\hat{\theta}_{LS} = Q^{-1}\mu$ (LS: least squares; solve the system, do not invert), $\hat{\theta}_{EN}$ (EN: elastic net) $= \arg\min_\theta \|\mu - Q\theta\|_2^2 + \lambda_2\|\theta\|_2^2 + \lambda_1\|\theta\|_1$ over a reported grid of penalty weights $\lambda_1, \lambda_2 \geq 0$, chosen by SR on 1987–1991.
  3. With weights fixed: SR of $F = \hat{\theta}^\top \tilde{F}$ on both test periods; paired SR differences on aligned months (HAC delta method or block bootstrap); the DSR over the grid.
  4. Loadings $\hat{\beta}_i = \overline{R_i^e F}$ on 1967–1986; project test returns on $\hat{\beta}$ each month; explained variation $1 - \overline{\hat{\epsilon}^2}/\overline{(R^e)^2}$, with $\hat{F}_t = \hat{\beta}^\top R_t^e/(\hat{\beta}^\top\hat{\beta})$ the monthly projection factor and $\hat{\epsilon}_{t,i} = R_{t,i}^e - \hat{\beta}_i\hat{F}_t$ its residual; the bar is a time average, and cross-sectional R^2 per characteristic; a chart of realised vs. implied mean returns.
- **Interpretation**
  - why can a linear no-arbitrage SDF beat a forecasting network (Chen, Pelger & Zhu (2024), Table I)?
  - does your post-2016 result survive the configurations tried?
- **Deliverables**: a Quantlet folder (code, `Metainfo.txt`, one chart) and one slide with the conclusion and its limitation

C3 Data Manifest: Kenneth French Deciles and Factors

- Download monthly factors from the Kenneth French Data Library: `F-F_Research_Data_Factors_CSV.zip` and `F-F_Research_Data_5_Factors_2x3_CSV.zip`.
- For the 15 characteristic sorts, save the exact archive names, decile columns, value/equal weighting, missing-value codes, units and common monthly dates in a manifest. Freeze the downloaded files before fitting.
- $\hat{\beta}_i$ (stage 4) is defined up to a constant; it cancels in $\hat{F}_t = \hat{\beta}^\top R_t^e / (\hat{\beta}^\top \hat{\beta})$. Compute the residuals with $\hat{F}_t$, not with F_t.

Acronyms used in this material (1/2)

- **ADF**: Augmented Dickey–Fuller (test)
- **AI**: Artificial Intelligence
- **AIC**: Akaike Information Criterion
- **AUC**: Area Under the (ROC) Curve
- **CI**: Confidence Interval
- **CV**: Cross-Validation
- **CW**: Clark–West (test for nested models)
- **DM**: Diebold–Mariano (test)
- **DSR**: Deflated Sharpe Ratio
- **ELW**: Exact Local Whittle (estimator)
- **EWMA**: Exponentially Weighted Moving Average
- **FF3**: Fama–French three-factor model
- **FF5**: Fama–French five-factor model
- **FFD**: Fixed-width window Fractional Differentiation
- **FFN**: FeedForward Network
- **FN**: False Negative
- **GAN**: Generative Adversarial Network
- **GB**: Gradient Boosting
- **GW**: Giacomini–White (test)
- **HAC**: Heteroskedasticity and Autocorrelation Consistent (standard errors)
- **IS**: In-Sample

Acronyms used in this material (2/2)

- **LASSO**: Least Absolute Shrinkage and Selection Operator
- **LS**: Linear SDF estimated by unregularised Least Squares
- **LW**: Local Whittle (estimator)
- **MDA**: Mean Decrease Accuracy
- **MDI**: Mean Decrease Impurity
- **OLS**: Ordinary Least Squares
- **OOS**: Out-Of-Sample
- **PBO**: Probability of Backtest Overfitting
- **PSR**: Probabilistic Sharpe Ratio
- **PT**: Pesaran–Timmermann (test of directional accuracy)
- **RC**: Reality Check (White's test for data snooping)
- **RF**: Random Forest
- **ROC**: Receiver Operating Characteristic (curve)
- **SD**: Standard Deviation
- **SDF**: Stochastic Discount Factor
- **SE**: Standard Error
- **SHAP**: SHapley Additive exPlanations
- **SPA**: Superior Predictive Ability (test)
- **SR**: Sharpe Ratio
- **TP**: True Positive
- **US**: United States
- **VIX**: CBOE Volatility IndeX

References (1/5)

- Bailey, D.H., Borwein, J., López de Prado, M., Zhu, Q.J. (2017). The probability of backtest overfitting. *Journal of Computational Finance*, 20(4), 39–69.
- Bailey, D.H., López de Prado, M. (2012). The Sharpe ratio efficient frontier. *Journal of Risk*, 15(2), 3–44.
- Bailey, D.H., López de Prado, M. (2014). The deflated Sharpe ratio: correcting for selection bias, backtest overfitting and non-normality. *Journal of Portfolio Management*, 40(5), 94–107.
- Belloni, A., Chernozhukov, V., Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2), 608–650.
- Bergmeir, C., Hyndman, R.J., Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120, 70–83.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Broadie, M., Glasserman, P., Kou, S. (1997). A continuity correction for discrete barrier options. *Mathematical Finance*, 7(4), 325–349.

References (2/5)

- Chen, L., Pelger, M., Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2), 714–750.
- Clark, T.E., West, K.D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291–311.
- DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3), 837–845.
- Dickey, D.A., Fuller, W.A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366), 427–431.
- Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Fama, E.F., French, K.R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22.
- Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
- Giacomini, R., White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6), 1545–1578.

References (3/5)

- Hanley, J.A., McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36.
- Hansen, P.R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4), 365–380.
- Hosking, J.R.M. (1981). Fractional differencing. *Biometrika*, 68(1), 165–176.
- Karatzas, I., Shreve, S.E. (1998). *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer.
- Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3), 1217–1241.
- Lo, A.W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal*, 58(4), 36–52.

References (4/5)

- López de Prado, M. (2018). *Advances in Financial Machine Learning*. Wiley.
- Lundberg, S.M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Newey, W.K., West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.
- Pesaran, M.H., Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics*, 10(4), 461–465.
- Politis, D.N., Romano, J.P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428), 1303–1313.
- Robinson, P.M. (1995). Gaussian semiparametric estimation of long range dependence. *Annals of Statistics*, 23(5), 1630–1661.
- Romano, J.P., Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), 1237–1282.

References (5/5)

- Shimotsu, K., Phillips, P.C.B. (2005). Exact local Whittle estimation of fractional integration. *Annals of Statistics*, 33(4), 1890–1933.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1), 267–288.
- White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097–1126.