Index

Page numbers followed by f or t refer to figure or table, respectively.

Absolute return attribution method, 68–69
Accounting data, 23, 169
Accuracy
  binary classification problems and, 52, 52f
  measurement of, 206
AdaBoost implementation, 100, 100f
Adaptable I/O System (ADIOS), 336–337, 339, 340
Alternative data, 24, 25
Amihud’s lambda, 288–289, 289f
Analytics, 24, 25
Annualized Sharpe ratio, 205
Annualized turnover, in backtesting, 196
Asset allocation
  classical areas of mathematics used in, 223–224
  covariance matrix in, 223, 224, 225f, 229, 231f, 232, 234
diversification in, 4, 9, 222, 224, 234, 238
Markowitz’s approach to, 221–222
Monte Carlo simulations for, 234–236, 235f–236f, 242–244
  numerical example of, 231–233, 232f, 233f, 233t
  practical problems in, 222–223, 223f
  quasi-diagonalization in, 224, 229, 233f
  recursive bisection in, 224, 229–231
  risk-based, 222. See also Risk-based asset allocation approaches
tree clustering approaches to,
  224–229, 225f, 228f, 232f
Attribution, 207–208
Augmented Dickey-Fuller (ADF) test, 253, 254, 256. See also Supremum augmented Dickey-Fuller (SADF) test
Average holding period, in backtesting, 196
Average slippage per turnover, 202
Backfilled data, 24, 152
Backtesters, 8
Backtesting, 139–244
  bet sizing in, 141–148
  common errors in, 151–157
Backtesting (Continued)
combinatorial purged cross-validation (CPCV) method in, 163–167
cross-validation (CV) for, 104, 162–163
customization of, 161
definition of, 151
“false discovery” probability and, 205
flawless completion as daunting task in, 152–153, 161
general recommendations on, 153–154
machine learning asset allocation and, 223–244
purpose of, 153
as research tool, 153, 154
strategy risk and, 211–218
strategy selection in, 155–156, 157f
synthetic data in, 169–192
uses of results of, 161
walk-forward (WF) method of, 161–162
Backtest overfitting, 4
backtesters’ evaluation of probability of, 8
bagging to reduce, 94–95, 100
combinatorial purged cross-validation (CPCV) method for, 166–167
concerns about risk of, 17, 119
cross-validation (CV) method and, 103, 155
decision trees and proneness to, 97
definition of, 153–154, 171
discretionary portfolio managers (PMs) and, 5
estimating extent of, 154
features stacking to reduce, 121–122
general recommendations on, 154
hyper-parameter tuning and, 129
need for skepticism, 11–12
optimal trading rule (OTR) framework for, 173–176
probability of. See Probability of backtest overfitting (PBO)
random forest (RF) method to reduce, 98, 99
selection bias and, 153–154
support vector machines (SVMs) and, 101
trading rules and, 171–172
walk-forward (WF) method and, 155, 162
Backtest statistics, 195–207
classification measurements in, 206–207
drawdown (DD) and time under water (TuW) in, 201, 202f
efficiency measurements in, 203–206
general description of, 196–197
holding period estimator in, 197
implementation shortfall in, 202–203
performance attribution and, 207–208
performance measurements in, 198
returns concentration in, 199–201
runs in, 199
run measurements in, 201–202
time-weighted rate of returns (TWRR) in, 198–199
timing of bets from series of target positions in, 197
types of, 195
Bagging, 94–97, 123
accuracy improvement using, 95–96, 97f
boosting compared with, 99–100
leakage reduction using, 105
observation redundancy and, 96–97
overfitting reduction and, 154
random forest (RF) method compared with, 98
scalability using, 101
variance reduction using, 94–95, 95f
Bars (table rows), 25–32
categories of, 26
dollar bars, 27–28, 28f, 44
dollar imbalance bars, 29–30
dollar runs bars, 31–32
INDEX

information-driven bars, 26, 29–32
standard bars, 26–28
tick bars, 26–27
tick imbalance bars, 29–30
tick runs bars, 31
time bars, 26, 43–44
volume bars, 27, 44
volume imbalance bars, 30–31
volume runs bars, 31–32
Becker-Parkinson volatility algorithm, 285–286
Bet sizing, 141–148
  average active bets approach in, 144
  bet concurrency calculation in, 141–142
  budgeting approach to, 142
  dynamic bet sizes and limit prices in, 145–148
  holding periods and, 144
  investment strategies and, 141
  meta-labeling approach to, 142
  performance attribution and, 207–208
  predicted probabilities and, 142–144, 143f
  runs and increase in, 199
  size discretization in, 144–145, 145f
  strategy-independent approaches to, 141–142
  strategy’s capacity and, 196
Bet timing, deriving, 197
Betting frequency
  backtesting and, 196
  computing, 215–216, 216f
  implied precision computation and, 214–215, 215f
  investment strategy with trade-off between precision and, 212–213, 212f
  strategy risk and, 211, 215
  targeting Sharpe ratio for, 212–213
  trade size and, 293
Bias, 93, 94, 100
Bid–ask spread estimator, 284–286
Bid wanted in competition (BWIC), 24, 286
big data analysis, 18, 236, 237f, 330, 331–332, 340
Bloomberg, 23, 36
Boosting, 99–100
  AdaBoost implementation of, 100, 100f
  bagging compared with, 99–100
  implementation of, 99
  main advantage of, 100
  variance and bias reduction using, 100
Bootstrap aggregation. See Bagging
Bootstraps, sequential, 63–66
Box-Jenkins analysis, 88
Broker fees per turnover, 202
Brown-Durbin-Evans CUSUM test, 250
Cancellation rates, 293–294
Capacity, in backtesting, 196
Chow-type Dickey-Fuller test, 251–252
Chu-Stinchcombe-White CUSUM test, 251
Classification models, 281–282
Classification problems
  class weights for underrepresented labels in, 71–72
  generating synthetic dataset for, 122
  meta-labeling and, 51–52, 142, 206–207
Classification statistics, 206–207
Class weights
  decision trees using, 99
  functionality for handling, 71–72
  underrepresented label correction using, 71
Cloud systems, 330–331, 334–335
Combinatorially symmetric cross-validation (CSCV) method, 155–156
Combinatorial purged cross-validation (CPCV) method, 163–167
algorithm steps in, 165
backtest overfitting and, 166–167
combinatorial splits in, 164–165, 164f
Combinatorial purged cross-validation (CPCV) method (Continued)
definition of, 163
examples of, 165–166
Compressed markets, 275
Computational Intelligence and Forecasting Technologies (CIFT) project, 329
Adaptable I/O System (ADIOS) and, 337
business applications developed by, 349–350
Flash Crash of 2010 response and, 341–343
mission of, 330, 331, 337
Conditional augmented Dickey-Fuller (CADF) test, 256, 256f, 257f
Correlation to underlying, in backtesting, 196
Corwin-Schultz algorithm, 284–286
Critical Line Algorithm (CLA), 221
description of, 222
Markowitz’s development of, 222
Monte Carlo simulations using,
234–236, 235f–236f, 242–244
numerical example with, 231–233,
232f, 233f, 233r
open-source implementation of, 222
practical problems with, 222–223,
223f
Cross-entropy loss (log loss) scoring,
133–134, 135f
Cross-validation (CV), 103–110
backtesting through, 104, 162–163
combinatorial purged cross-validation (CPCV) method in, 163–167
embargo on training observations in,
107–108, 108f
failures in finance using, 104
goal of, 103
hyper-parameter tuning with,
129–135
k-fold, 103–109, 104f
leakage in, 104–105
model development and, 104

overlapping training observations in,
109
purpose of, 103
purging process in training set for
leakage reduction in, 105–106,
107f
sklearn bugs in, 109–110
CUSUM filter, 38–40, 40f
CUSUM tests, 249, 250–251
CV. See Cross-validation

Data analysis, 21–90
financial data structures and, 23–40
fractionally differentiated features
and, 75–88
labeling and, 43–55
sample weights and, 59–72
Data curators, 7
Data mining and data snooping, 152
Decision trees, 97–99
Decompressed markets, 275
Deflated Sharpe ratio (DSR), 204, 205f
Deployment team, 8
Dickey-Fuller test
Chow type, 251–252
supremum augmented (SADF),
252–259, 253f, 257f
Discretionary portfolio managers (PMs), 4–5, 15
Discretization of bet size, 144–145,
145f
Diversification, 4, 9, 222, 224, 234,
238
Dollar bars, 27–28, 28f, 44
Dollar imbalance bars (DIBs), 29–30
Dollar performance per turnover, 202
Dollar runs bars (DRBs), 31–32
Downsampling, 38
Drawdown (DD)
definition of, 201
deriving, 201
example of, 202f
run measurements using, 202
Dynamic bet sizes, 145–148
INDEX

Econometrics, 14, 85
financial Big Data analysis and, 236
financial machine learning versus, 15
HRP approach compared with, 236, 237
investment strategies based in, 6
paradigms used in, 88
substitution effects and, 114
trading rules and, 169
Efficiency measurements, 203–206
annualized Sharpe ratio and, 205
deflated Sharpe ratio (DSR) and, 204, 205
information ratio and, 205
probabilistic Sharpe ratio (PSR) and, 203–204, 204f, 205–206, 218
Sharpe ratio (SR) definition in, 203
Efficient frontier, 222
Electricity consumption analysis, 340–341, 342f–343f
Engle-Granger analysis, 88
Ensemble methods, 93–101
boosting and, 99–100
bootstrap aggregation (bagging) and, 94–97, 101
random forest (RF) method and, 97–99
Entropy features, 263–277
encoding schemes in estimates of, 269–271
financial applications of, 275–277
generalized mean and, 271–275, 274f
Lempel-Ziv (LZ) estimator in, 265–269
maximum likelihood estimator in, 264–265
Shannon’s approach to, 263–264
ETF trick, 33–34, 84, 253
Event-based sampling, 38–40, 40f
Excess returns, in information ratio, 205
Execution costs, 202–203
Expanding window method, 80–82, 81f
Explosiveness tests, 249, 251–259
Chow-type Dickey-Fuller test, 251–252
supremum augmented Dickey-Fuller (SADF) test, 252–259, 253f, 257
Factory plan, 5, 11
Feature analysts, 7
Feature importance, 113–127
features stacking approach to, 121–122
importance of, 113–114
mean decrease accuracy (MDA) and, 116–117
mean decrease impurity (MDI) and, 114–116
orthogonal features and, 118–119
parallelized approach to, 121
plotting function for, 124–125
random forest (RF) method and, 98
as research tool, 153
single feature importance (SFI) and, 117–118
synthetic data experiments with, 122–124
weighted Kendall’s tau computation in, 120–121
without substitution effects, 117–121
with substitution effects, 114–117
Features stacking importance, 121–122
Filter trading strategy, 39
Finance
algorithmization of, 14
human investors’ abilities in, 4, 14
purpose and role of, 4
usefulness of ML algorithms in, 4, 14
Financial data
alternative, 25
analytics and, 25
essential types of, 23, 24f
fundamental, 23–24
market, 24–25
Financial data structures, 23–40
bars (table rows) in, 25–32
INDEX

Financial data structures (Continued)
- multi-product series in, 32–37
- sampling features in, 38–40
- unstructured, raw data as starting point for, 23

Financial Information eXchange (FIX) messages, 7, 24, 25, 281

Financial machine learning econometrics versus, 15
- prejudices about use of, 16
- standard machine learning separate from, 4

Financial machine learning projects reasons for failure of, 4–5
- structure by strategy component in, 9–12
- Fixed-time horizon labeling method, 43–44
- Fixed-width window fracdiff (FFD) method, 82–84, 83f

Flash crashes, 296
- class weights for predicting, 71
- “early warning” indicators in, 345
- high performance computing (HPC) tools and, 347–348
- signs of emerging illiquidity events and, 331

F1 scores
- measurement of, 206
- meta-labeling and, 52–53

Fractional differentiation
- data transformation method for stationarity in, 77–78
- earlier methods of, 76–77
- expanding window method for, 80–82, 81f
- fixed-width window fracdiff (FFD) method for, 82–84, 83f
- maximum memory preservation in, 84–85, 84f, 86r–87r

Frequency. See Betting frequency

Fundamental data, 23–24, 24t
- Fusion collaboration project, 338–340, 339f

Futures
dollar bars and, 28
- ETF trick with, 33–34
- non-negative rolled price series and, 37
- single futures roll method with, 36–37
- volume bars and, 27

Gaps series, in single future roll method, 36–37

Global Investment Performance Standards (GIPS), 161, 195, 198

Graph theory, 221, 224
- Grid search cross-validation, 129–131

Hasbrouck’s lambda, 289, 290f

Hedging weights, 35

Herfindahl-Hirschman Index (HHI) concentration, 200, 201
- HHI indexes
  - on negative returns, 202
  - on positive returns, 201
  - on time between bets, 202
- Hierarchical Data Format 5 (HDF5), 336
- Hierarchical Risk Parity (HRP) approach
  - econometric regression compared with, 236, 237f
  - full implementation of, 240–242
  - Monte Carlo simulations using, 234–236, 235f–236f, 242–244
  - numerical example of, 231–233, 232f, 233f, 233r
  - practical application of, 221
  - quasi-diagonalization in, 224, 229
  - recursive bisection in, 224, 229–231
  - standard approaches compared with, 221, 236–238
INDEX

traditional risk parity approach compared with, 231–232, 233, 234, 235

tree clustering approaches to, 224–229, 225f, 228f, 232

High-frequency trading, 14, 196, 212

High-low volatility estimator, 283–284

High-performance computing (HPC), 301–349

ADIOS and, 336–337, 339, 340

atoms and molecules in parallelization and, 306–309

CIFT business applications and, 349–350

cloud systems compared with, 334–335

combinatorial optimization and, 319–320

electricity consumption analysis using, 340–341, 342–343


fusion collaboration project using, 338–340, 339

global dynamic optimal trajectory and, 325–327

hardware for, 331–335, 332, 333, 334

integer optimization approach and, 321–325, 322

multiprocessing and, 304–306, 309–316

objective function and, 320–321

pattern-finding capability in, 330–331

software for, 335–337

streaming data analysis using, 329

supernova hunting using, 337–338, 338

use cases for, 337–349

vectorization and, 303–304

Holding periods

backtesting and, 196

bet sizing and, 144

estimating in strategy, 196, 197

optimal trading rule (OTR) algorithm with, 174, 175

triple-period labeling method and, 46

Hyper-parameter tuning, 129–135

grid search cross-validation and, 129–131

log loss scoring used with, 133–134, 135

randomized search cross-validation and, 131–133

Implementation shortfall statistics, 202–203

Implied betting frequency, 215–216, 216

Implied precision computation, 214–215, 215

Indicator matrix, 64–65, 66, 67

Information-driven bars (table rows), 26, 29–32

dollar imbalance bars, 29–30

dollar runs bars, 31–32

purpose of, 29

tick imbalance bars, 29–30

tick runs bars, 31

volume imbalance bars, 30–31

volume runs bars, 31–32

Information ratio, 205

Inverse-Variance Portfolio (IVP)

asset allocation numerical example of, 231–233, 232f, 233f, 233

Monte Carlo simulations using, 234–236, 235f–236f, 242–244

Investment strategies

algorithmization of, 14

bet sizing in, 141

evolution of, 6

exit conditions in, 211

human investors’ abilities and, 4, 14

log loss scoring used with hyper-parameter tuning in, 134–135, 135
Investment strategies (Continued)
profit-taking and stop-loss limits in, 170–171, 172, 211
risk in. See Strategy risk
structural breaks and, 249
trade-off between precision and frequency in, 212–213, 212f
trading rules and algorithms in, 169–170
Investment strategy failure probability, 216–218
algorithm in, 217
implementation of algorithm in, 217–218
probabilistic Sharpe ratio (PSR) similarity to, 218
strategy risk and, 216–217

K-fold cross-validation (CV), 103–109
description of, 103–104, 104f
embargo on training observations in, 107–108, 108f
leakage in, 104–105
mean decrease accuracy (MDA) feature with, 116
overlapping training observations in, 109
purging process in training set for leakage reduction in, 105–106, 107f
when used, 104
Kyle’s lambda, 286–288, 288f

Labeling, 43–55
daily volatility at intraday estimation for, 44–45
dropping unnecessary or under-populated labels in, 54–55
fixed-time horizon labeling method for, 43–44
learning side and size in, 48–50
meta-labeling and, 50–53
quantamental approach using, 53–54
triple-barrier labeling method for, 45–46, 47f
Labels
average uniqueness over lifespan of, 61–62, 61f
class weights for underrepresented labels, 71–72
estimating uniqueness of, 60–61
Lawrence Berkeley National Laboratory (LBNL, Berkeley Lab), 18, 329, 331
Leakage, and cross-validation (CV), 104–105
Leakage reduction
bagging for, 105
purging process in training set for, 105–106, 107f
sequential bootstraps for, 105
walk-forward timefolds method for, 155
Lempel-Ziv (LZ) estimator, 265–269
Leverage, in backtesting, 196
Limit prices, in bet sizing, 145–148
Log loss scoring, in hyper-parameter tuning, 133–134, 135f
Log-uniform distribution, 132–133
Look-ahead bias, 152

Machine learning (ML), 3
finance and, 4, 14
financial machine learning separate from, 4
HRP approach using, 221, 224
human investors and, 4, 14
prejudices about use of, 16
Machine learning asset allocation,
223–244. See also Hierarchical Risk Parity (HRP) approach
Monte Carlo simulations for,
234–236, 235f–236f, 242–244
numerical example of, 231–233, 232f, 233f, 233r
quasi-diagonalization in, 224, 229, 233f
INDEX

recursive bisection in, 224, 229–231
tree clustering approaches to, 224–229, 225f, 228f, 232f
Market data, 24–25, 24r
Markowitz, Harry, 221–222
Maximum dollar position size, in backtesting, 196
Maximum likelihood estimator, in entropy, 264–265
Mean decrease accuracy (MDA) feature importance, 116–117
computed on synthetic dataset, 125–126, 126f
considerations in working with, 116
implementation of, 116–117
single feature importance (SFI) and, 127
Mean decrease impurity (MDI) feature importance, 114–116
computed on synthetic dataset, 125, 125f
considerations in working with, 115
implementation of, 115–116
single feature importance (SFI) and, 127
Message Passing Interface (MPI), 335
Meta-labeling, 50–55
bet sizing using, 142
code enhancements for, 50–51
description of, 50, 127
dropping unnecessary or under-populated labels in, 54–55
how to use, 51–53
quantamental approach using, 53–54
Meta-strategy paradigm, 5, 6, 11, 18
Microstructural features, 281–296
alternative features of, 293–295
Amihud’s lambda and, 288–289, 289f
bid-ask spread estimator (Corwin-Schultz algorithm) and, 284–286
Hasbrouck’s lambda and, 289, 290f
high-low volatility estimator and, 283–284
Kyle’s lambda and, 286–288, 288f
microstructural information definition and, 295–296
probability of informed trading and, 290–292
Roll model and, 282–283
sequential trade models and, 290
strategic trade models and, 286
tick rule and, 282
volume-synchronized probability of informed trading (VPIN) and, 292
Mixture of Gaussians (EF3M), 141–142, 149, 217–219
Model development
cross-validation (CV) for, 104, 108–109
overfitting reduction and, 154
single feature importance method and, 117
Modelling, 91–135
applications of entropy to, 275
backtesting in, 153
cross-validation in, 103–110
econometrics and, 15
ensemble methods in, 93–101
entropy features in, 275–277
feature importance in, 113–127
hyper-parameter tuning with cross-validation in, 129–135
market microstructure theories and, 281–282
three sources of errors in, 93
Monte Carlo simulations
machine learning asset allocation and, 234–236, 235f–236f, 242–244
sequential bootstraps evaluation using, 66–68, 68f
Multi-product series, 32–37
ETF trick for, 33–34
PCA weights for, 35–36, 35f
single future roll in, 36–37
National laboratories, 5, 10, 18
Negative (neg) log loss scores
Negative (neg) log loss scores (Continued)
hyper-parameter tuning using, 133–134, 135f
measurement of, 207
Noise, causes of, 93
Non-negative rolled price series, 37

Optimal trading rule (OTR) framework, 173–176
algorithm steps in, 173–174
cases with negative long-run equilibrium in, 182–187, 186f, 187f–191f
cases with positive long-run equilibrium in, 180–182, 181f, 182f, 183f–186f
cases with zero long-run equilibrium in, 177–180, 177f, 178f, 179f
experimental results using simulation in, 176–191
implementation of, 174–176
overfitting and, 172
profit-taking and stop-loss limits in, 173–208, 176–177, 192
synthetic data for determination of, 192
Options markets, 295
Ornstein-Uhlenbeck (O-U) process, 172–173
Orthogonal features, 118–119
benefits of, 119
computation of, 119
implementation of, 118–119
Outliers, in quantitative investing, 152
Overfitting. See Backtest overfitting

Parallelized feature importance, 121
PCA (see Principal components analysis)
Performance attribution, 207–208
Plotting function for feature importance, 124–125

PnL (mark-to-market profits and losses)
ETF trick and, 33
performance attribution using, 207–208
as performance measurement, 198
rolled prices for simulating, 37
PnL from long positions, 198
Portfolio construction. See also
Hierarchical Risk Parity (HRP) approach
classical areas of mathematics used in, 223–224
covariance matrix in, 223, 224, 225f, 229, 231f, 232, 234
diversification in, 4, 9, 222, 224, 234, 238
entropy and concentration in, 275–276
Markowitz’s approach to, 221–222
Monte Carlo simulations for, 234–236, 235f–236f, 242–244
numerical example of, 231–233, 232f, 233f, 233t
practical problems in, 222–223, 223f
tree clustering approaches to, 224–229, 225f, 228f, 232f
Portfolio oversight, 8–9
Portfolio risk. See also Hierarchical Risk Parity (HRP) approach; Risk; Strategy risk
portfolio decisions based on, 221–222
probability of strategy failure and, 217
strategy risk differentiated from, 217
Portfolio turnover costs, 202–203
Precision
binary classification problems and, 52–53, 52f
investment strategy with trade-off between frequency and, 212–213, 212f
measurement of, 206
Predicted probabilities, in bet sizing, 142–144, 143f
INDEX

Principal components analysis (PCA) hedging weights using, 35–36, 35f
linear substitution effects and, 118, 119–121

Probabilistic Sharpe ratio (PSR) calculation of, 203–204, 204f
as efficiency statistic, 203, 205–206
probability of strategy failure, similarity to, 218

Probability of backtest overfitting (PBO) backtest overfitting evaluation using, 171–172
combinatorially symmetric cross-validation (CSCV) method for, 155–156
strategy selection based on estimation of, 155–156, 157f, 171

Probability of informed trading (PIN), 276, 290–292

Probability of strategy failure, 216–218
algorithm in, 217
implementation of algorithm in, 217–218
probabilistic Sharpe ratio (PSR), similarity to, 218
strategy risk and, 211–217

Profit-taking, and investment strategy exit, 211

Profit-taking limits asymmetric payoff dilemma and, 178–180
cases with negative long-run equilibrium and, 182–187, 186f, 187f–191f
cases with positive long-run equilibrium and, 180–182, 181f, 182f, 183f–186f
cases with zero long-run equilibrium and, 177–180, 177f, 178f, 179f
daily volatility at intraday estimation points computation and, 44–45
investment strategies using, 170–171, 172
learning side and size and, 48

optimal trading rule (OTR) algorithm for, 173–174, 176–177, 192
strategy risk and, 211

Purged K-fold cross-validation (CV) grid search cross-validation and, 129–131
hyper-parameter tuning with, 129–135
implementation of, 105–106, 107f
randomized search cross-validation and, 131–133
Python, 14

Quantamental approach, 4, 15, 53–54
Quantamental funds, 19

Quantitative investing backtest overfitting in, 113, 154
failure rate in, 4
meta-strategy paradigm in, 5
quantamental approach in, 53
seven sins of, 152, 153

Quantum computing, 319–328

Random forest (RF) method, 97–99
alternative ways of setting up, 98–99
bagging compared with, 98
bet sizing using, 142
Randomized search cross-validation, 131–133
Recall binary classification problems and, 52, 52f
measurement of, 206
Reinstated value, 24
Return attribution method, 68–69
Return on execution costs, 203

Returns concentration, 199–201
RF. See Random forest (RF) method
Right-tail unit-root tests, 250
Risk. See also Hierarchical Risk Parity (HRP) approach; Strategy risk backtest statistics for uncovering, 195 entropy application to portfolio concentration and, 276 liquidity and, 7, 286 ML algorithms for monitoring, 14 PCA weights and, 35–36, 35f portfolio oversight and, 8 profit taking and stop-loss limits and, 44 structural breaks and, 249 walk-forward (WF) approach and, 161 Risk-based asset allocation approaches, 222 HRP approach comparisons in, 236–238 structural breaks and, 249 Risk parity, 222. See also Hierarchical Risk Parity (HRP) approach HRP approach compared with traditional approach to, 231–232, 233r, 234, 235f Rolled prices, 37 Roll model, 282–283 Sample weights, 59–72 average uniqueness of labels over lifespan and, 61–62, 61f bagging classifiers and uniqueness and, 62–63 indicator matrix for, 64–65 mean decrease accuracy (MDA) feature importance with, 116 number of concurrent labels and, 60–61 overlapping outcomes and, 59–60 return attribution method and, 68–69 sequential bootstrap and, 63–68 time-decay factors and, 70–71, 72f Sampling features, 38–40 downsampling and, 38 event-based sampling and, 38–40 Scalability bagging for, 101 sample size in ML algorithms and, 38, 101 Scikit-learn (sklearn) class weights in, 71 cross-validation (CV) bugs in, 109–110 grid search cross-validation in, 129–130 labels and bug in, 55, 72, 94 mean decrease impurity (MDI) and, 115 neg log loss as scoring statistic and bug in, 134 observation redundancy and bagging classifiers in, 97 random forest (RF) overfitting and, 98–99 support vector machine (SVM) implementation in, 101 synthetic dataset generation in, 122 walk-forward timefolds method in, 155 Selection bias, 153–154, 167 Sequential bootstraps, 63–68 description of, 63–64 implementation of, 64–65 leakage reduction using, 105 Monte Carlo experiments evaluating, 66–68, 68f numerical example of, 65–66 Shannon, Claude, 263–264 Sharpe ratio (SR) in efficiency measurements annualized, 205 definition of, 203 deflated (DSR), 204, 205f information ratio and, 205 probabilistic (PSR), 203–204, 204f, 205–206, 218 purpose of, 203 targeting, for various betting frequencies, 212–213 Shorting, in quantitative investing, 152
### INDEX

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal order flows</td>
<td>295</td>
</tr>
<tr>
<td>Simulations, overfitting of</td>
<td>154</td>
</tr>
<tr>
<td>Single feature importance (SFI),</td>
<td>117–118, 125–127, 126f</td>
</tr>
<tr>
<td>Single future roll</td>
<td>36–37</td>
</tr>
<tr>
<td>Sklearn. See Scikit-learn</td>
<td></td>
</tr>
<tr>
<td>Stacked feature importance,</td>
<td>121–122</td>
</tr>
<tr>
<td>Standard bars (table rows),</td>
<td>26–28</td>
</tr>
<tr>
<td>dollar bars, 27–28, 28f, 44</td>
<td></td>
</tr>
<tr>
<td>purpose of, 26</td>
<td></td>
</tr>
<tr>
<td>tick bars, 26–27</td>
<td></td>
</tr>
<tr>
<td>time bars, 26, 43–44</td>
<td></td>
</tr>
<tr>
<td>volume bars, 27, 44</td>
<td></td>
</tr>
<tr>
<td>Stationarity</td>
<td></td>
</tr>
<tr>
<td>data transformation method to ensure,</td>
<td>77–78</td>
</tr>
<tr>
<td>fractional differentiation applied to,</td>
<td>76–77</td>
</tr>
<tr>
<td>fractional differentiation implementation methods for,</td>
<td>80–84</td>
</tr>
<tr>
<td>integer transformation for,</td>
<td>76</td>
</tr>
<tr>
<td>maximum memory preservation for, 84–85, 84f, 86f–87f</td>
<td></td>
</tr>
<tr>
<td>memory loss dilemma and,</td>
<td>75–76</td>
</tr>
<tr>
<td>Stop-loss, and investment strategy exit,</td>
<td>211</td>
</tr>
<tr>
<td>Stop-loss limits</td>
<td></td>
</tr>
<tr>
<td>asymmetric payoff dilemma and, 178–180</td>
<td></td>
</tr>
<tr>
<td>cases with negative long-run equilibrium and, 182–187, 186f, 187f–191f</td>
<td></td>
</tr>
<tr>
<td>cases with positive long-run equilibrium and, 180–182, 181f, 182f, 183f–186f</td>
<td></td>
</tr>
<tr>
<td>cases with zero long-run equilibrium and, 177–180, 177f, 178f, 179f</td>
<td></td>
</tr>
<tr>
<td>daily volatility computation and, 44–45</td>
<td></td>
</tr>
<tr>
<td>fixed-time horizon labeling method and, 44</td>
<td></td>
</tr>
<tr>
<td>investment strategies using, 170–171, 172, 211</td>
<td></td>
</tr>
<tr>
<td>learning side and size and,</td>
<td>48</td>
</tr>
<tr>
<td>optimal trading rule (OTR) algorithm for, 173–174, 176–177, 192</td>
<td></td>
</tr>
<tr>
<td>strategy risk and, 211</td>
<td></td>
</tr>
<tr>
<td>triple-barrier labeling method for, 45–46, 47f</td>
<td></td>
</tr>
<tr>
<td>Storytelling, 162</td>
<td></td>
</tr>
<tr>
<td>Strategists, 7</td>
<td></td>
</tr>
<tr>
<td>Strategy risk, 211–218</td>
<td></td>
</tr>
<tr>
<td>asymmetric payouts and, 213–216</td>
<td></td>
</tr>
<tr>
<td>calculating, 217, 218</td>
<td></td>
</tr>
<tr>
<td>implied betting frequency and, 215–216, 216f</td>
<td></td>
</tr>
<tr>
<td>implied precision and, 214–215, 215f</td>
<td></td>
</tr>
<tr>
<td>investment strategies and understanding of, 211</td>
<td></td>
</tr>
<tr>
<td>portfolio risk differentiated from, 217</td>
<td></td>
</tr>
<tr>
<td>probabilistic Sharpe ratio (PSR) similarity to, 218</td>
<td></td>
</tr>
<tr>
<td>strategy failure probability and, 216–218</td>
<td></td>
</tr>
<tr>
<td>symmetric payouts and, 211–213, 212f</td>
<td></td>
</tr>
<tr>
<td>Structural breaks, 249–261</td>
<td></td>
</tr>
<tr>
<td>CUSUM tests in, 250–251</td>
<td></td>
</tr>
<tr>
<td>explosiveness tests in, 249, 251–259</td>
<td></td>
</tr>
<tr>
<td>sub- and super-martingale tests in, 259–261</td>
<td></td>
</tr>
<tr>
<td>types of tests in, 249–250</td>
<td></td>
</tr>
<tr>
<td>Sub- and super-martingale tests, 250, 259–261</td>
<td></td>
</tr>
<tr>
<td>Supernova research, 337–338, 338f</td>
<td></td>
</tr>
<tr>
<td>Support vector machines (SVMs), 38, 101</td>
<td></td>
</tr>
<tr>
<td>Supremum augmented Dickey-Fuller (SADF) test, 252–259, 253f, 257f</td>
<td></td>
</tr>
<tr>
<td>conditional ADF, 256, 256f, 257f</td>
<td></td>
</tr>
<tr>
<td>implementation of, 258–259</td>
<td></td>
</tr>
<tr>
<td>quantile ADF, 255–256</td>
<td></td>
</tr>
<tr>
<td>Survivorship bias, 152</td>
<td></td>
</tr>
<tr>
<td>SymPy Live, 214</td>
<td></td>
</tr>
<tr>
<td>Synthetic data</td>
<td></td>
</tr>
<tr>
<td>backtesting using, 169–192</td>
<td></td>
</tr>
</tbody>
</table>
Synthetic data (Continued)
experimental results using simulation combinations with, 176–191
optimal trading rule (OTR) framework using, 173–176

Tick bars, 26–27
Tick imbalance bars (TIBs), 29–30
Tick rule, 282
Tick runs bars (TRBs), 31

Time bars
description of, 26
fixed-time horizon labeling method using, 43–44

Time-decay factors, and sample weights, 70–71, 72f

Time period, in backtesting, 196

Time series
fractional differentiation applied to, 76
integer transformation for stationarity in, 76
stationarity vs. memory loss dilemma in, 75–76

Time under water (TuW)
definition of, 201
deriving, 201
equation of, 202f
run measurements using, 202

Time-weighted average price (TWAP), 24, 294

Time-weighted rate of returns (TWRR), 198–199

Trading rules
investment strategies and algorithms in, 169–170

optimal trading rule (OTR) framework for, 173–176
overfitting in, 171–172
Transaction costs, in quantitative investing, 152
Tree clustering approaches, in asset allocation, 224–229, 225f, 228f, 232f
Triple-barrier labeling method, 45–46, 47f, 145
Turnover costs, 202–203

Variance
boosting to reduce, 100
causes of, 93
ensemble methods to reduce, 94
random forest (RF) method for, 97–98
Vectorization, 303–304

Volume bars, 27, 44
Volume imbalance bars (VIBs), 30–31
Volume runs bars (VRBs), 31–32
Volume-synchronized probability of informed trading (VPIN), 276, 282, 292

Walk-forward (WF) method
backtesting using, 161–162
overfitting in, 155, 162
pitfalls of, 162
Sharpe ratio estimation in, 166
two key advantages of, 161–162
Walk-forward timefolds method, 155
Weighted Kendall’s tau, 120–121
Weights. See Class weights; Sample weights