Advances in Financial Machine Learning

MARCOS LOPEZ DE PRADO

WILEY

Contents

About the Author **xxi**

PREAMBLE **1**

1 Financial Machine Learning as a Distinct Subject **3**

1.1 Motivation, 3

1.2 The Main Reason Financial Machine Learning Projects Usually Fail, 4

1.2.1 The Sisyphus Paradigm, 4

1.2.2 The Meta-Strategy Paradigm, 5

1.3 Book Structure, 6

1.3.1 Structure by Production Chain, 6

1.3.2 Structure by Strategy Component, 9

1.3.3 Structure by Common Pitfall, 12

1.4 Target Audience, 12

1.5 Requisites, 13

1.6 FAQs, 14

1.7 Acknowledgments, 18

Exercises, 19

References, 20

Bibliography, 20

PART 1 DATA ANALYSIS **21**

2 Financial Data Structures **23**

2.1 Motivation, 23

2.2 Essential Types of Financial Data, 23
    2.2.1 Fundamental Data, 23
    2.2.2 Market Data, 24
    2.2.3 Analytics, 25
    2.2.4 Alternative Data, 25
2.3 Bars, 25
    2.3.1 Standard Bars, 26
    2.3.2 Information-Driven Bars, 29
2.4 Dealing with Multi-Product Series, 32
    2.4.1 The ETF Trick, 33
    2.4.2 PCA Weights, 35
    2.4.3 Single Future Roll, 36
2.5 Sampling Features, 38
    2.5.1 Sampling for Reduction, 38
    2.5.2 Event-Based Sampling, 38
Exercises, 40
References, 41

Labeling

3.1 Motivation, 43
3.2 The Fixed-Time Horizon Method, 43
3.3 Computing Dynamic Thresholds, 44
3.4 The Triple-Barrier Method, 45
3.5 Learning Side and Size, 48
3.6 Meta-Labeling, 50
3.7 How to Use Meta-Labeling, 51
3.8 The Quantamental Way, 53
3.9 Dropping Unnecessary Labels, 54
Exercises, 55
Bibliography, 56

Sample Weights

4.1 Motivation, 59
4.2 Overlapping Outcomes, 59
4.3 Number of Concurrent Labels, 60
4.4 Average Uniqueness of a Label, 61
4.5 Bagging Classifiers and Uniqueness, 62
    4.5.1 Sequential Bootstrap, 63
    4.5.2 Implementation of Sequential Bootstrap, 64

4.5.3 A Numerical Example, 65
4.5.4 Monte Carlo Experiments, 66
4.6 Return Attribution, 68
4.7 Time Decay, 70
4.8 Class Weights, 71
Exercises, 72
References, 73
Bibliography, 73

5 Fractionally Differentiated Features 75

5.1 Motivation, 75
5.2 The Stationarity vs. Memory Dilemma, 75
5.3 Literature Review, 76
5.4 The Method, 77
5.4.1 Long Memory, 77
5.4.2 Iterative Estimation, 78
5.4.3 Convergence, 80
5.5 Implementation, 80
5.5.1 Expanding Window, 80
5.5.2 Fixed-Width Window Fracdiff, 82
5.6 Stationarity with Maximum Memory Preservation, 84
5.7 Conclusion, 88
Exercises, 88
References, 89
Bibliography, 89

PART 2 MODELLING 91

6 Ensemble Methods 93

6.1 Motivation, 93
6.2 The Three Sources of Errors, 93
6.3 Bootstrap Aggregation, 94
6.3.1 Variance Reduction, 94
6.3.2 Improved Accuracy, 96
6.3.3 Observation Redundancy, 97
6.4 Random Forest, 98
6.5 Boosting, 99

6.6 Bagging vs. Boosting in Finance, 100
6.7 Bagging for Scalability, 101
Exercises, 101
References, 102
Bibliography, 102

7 Cross-Validation in Finance

7.1 Motivation, 103
7.2 The Goal of Cross-Validation, 103
7.3 Why K-Fold CV Fails in Finance, 104
7.4 A Solution: Purged K-Fold CV, 105
7.4.1 Purging the Training Set, 105
7.4.2 Embargo, 107
7.4.3 The Purged K-Fold Class, 108
7.5 Bugs in Sklearn's Cross-Validation, 109
Exercises, 110
Bibliography, 111

8 Feature Importance

8.1 Motivation, 113
8.2 The Importance of Feature Importance, 113
8.3 Feature Importance with Substitution Effects, 114
8.3.1 Mean Decrease Impurity, 114
8.3.2 Mean Decrease Accuracy, 116
8.4 Feature Importance without Substitution Effects, 117
8.4.1 Single Feature Importance, 117
8.4.2 Orthogonal Features, 118
8.5 Parallelized vs. Stacked Feature Importance, 121
8.6 Experiments with Synthetic Data, 122
Exercises, 127
References, 127

9 Hyper-Parameter T\ining with Cross-Validation

9.1 Motivation, 129
9.2 Grid Search Cross-Validation, 129
9.3 Randomized Search Cross-Validation, 131
9.3.1 Log-Uniform Distribution, 132
9.4 Scoring and Hyper-parameter Tuning, 134

Exercises, 135
References, 136
Bibliography, 137

PART 3 BACKTESTING 139

10 Bet Sizing 141

10.1 Motivation, 141
10.2 Strategy-Independent Bet Sizing Approaches, 141
10.3 Bet Sizing from Predicted Probabilities, 142
10.4 Averaging Active Bets, 144
10.5 Size Discretization, 144
10.6 Dynamic Bet Sizes and Limit Prices, 145
Exercises, 148
References, 149
Bibliography, 149

11 The Dangers of Backtesting 151

11.1 Motivation, 151
11.2 Mission Impossible: The Flawless Backtest, 151
11.3 Even If Your Backtest Is Flawless, It Is Probably Wrong, 152
11.4 Backtesting Is Not a Research Tool, 153
11.5 A Few General Recommendations, 153
11.6 Strategy Selection, 155
Exercises, 158
References, 158
Bibliography, 159

12 Backtesting through Cross-Validation 161

12.1 Motivation, 161
12.2 The Walk-Forward Method, 161
12.2.1 Pitfalls of the Walk-Forward Method, 162
12.3 The Cross-Validation Method, 162
12.4 The Combinatorial Purged Cross-Validation Method, 163
12.4.1 Combinatorial Splits, 164
12.4.2 The Combinatorial Purged Cross-Validation Backtesting Algorithm, 165
12.4.3 A Few Examples, 165

12.5 How Combinatorial Purged Cross-Validation Addresses Backtest Overfitting, 166

Exercises, 167

References, 168

13 Backtesting on Synthetic Data

13.1 Motivation, 169

13.2 Trading Rules, 169

13.3 The Problem, 170

13.4 Our Framework, 172

13.5 Numerical Determination of Optimal Trading Rules, 173

13.5.1 The Algorithm, 173

13.5.2 Implementation, 174

13.6 Experimental Results, 176

13.6.1 Cases with Zero Long-Run Equilibrium, 177

13.6.2 Cases with Positive Long-Run Equilibrium, 180

13.6.3 Cases with Negative Long-Run Equilibrium, 182

13.7 Conclusion, 192

Exercises, 192

References, 193

14 Backtest Statistics

14.1 Motivation, 195

14.2 Types of Backtest Statistics, 195

14.3 General Characteristics, 196

14.4 Performance, 198

14.4.1 Time-Weighted Rate of Return, 198

14.5 Runs, 199

14.5.1 Returns Concentration, 199

14.5.2 Drawdown and Time under Water, 201

14.5.3 Runs Statistics for Performance Evaluation, 201

14.6 Implementation Shortfall, 202

14.7 Efficiency, 203

14.7.1 The Sharpe Ratio, 203

14.7.2 The Probabilistic Sharpe Ratio, 203

14.7.3 The Deflated Sharpe Ratio, 204

14.7.4 Efficiency Statistics, 205

14.8 Classification Scores, 206

14.9 Attribution, 207

Exercises, 208
References, 209
Bibliography, 209

15 Understanding Strategy Risk 211

15.1 Motivation, 211
15.2 Symmetric Payouts, 211
15.3 Asymmetric Payouts, 213
15.4 The Probability of Strategy Failure, 216
15.4.1 Algorithm, 217
15.4.2 Implementation, 217
Exercises, 219
References, 220

16 Machine Learning Asset Allocation 221

16.1 Motivation, 221
16.2 The Problem with Convex Portfolio Optimization, 221
16.3 Markowitz's Curse, 222
16.4 From Geometric to Hierarchical Relationships, 223
16.4.1 Tree Clustering, 224
16.4.2 Quasi-Diagonalization, 229
16.4.3 Recursive Bisection, 229
16.5 A Numerical Example, 231
16.6 Out-of-Sample Monte Carlo Simulations, 234
16.7 Further Research, 236
16.8 Conclusion, 238
Appendices, 239
16.A.1 Correlation-based Metric, 239
16.A.2 Inverse Variance Allocation, 239
16.A.3 Reproducing the Numerical Example, 240
16.A.4 Reproducing the Monte Carlo Experiment, 242
Exercises, 244
References, 245

PART 4 USEFUL FINANCIAL FEATURES 247

17 Structural Breaks 249

17.1 Motivation, 249
17.2 Types of Structural Break Tests, 249

17.3 CUSUM Tests, 250

17.3.1 Brown-Durbin-Evans CUSUM Test on Recursive Residuals, 250

17.3.2 Chu-Stinchcombe-White CUSUM Test on Levels, 251

17.4 Explosiveness Tests, 251

17.4.1 Chow-Type Dickey-Fuller Test, 251

17.4.2 Supremum Augmented Dickey-Fuller, 252

17.4.3 Sub-and Super-Martingale Tests, 259

Exercises, 261

References, 261

18 Entropy Features 263

18.1 Motivation, 263

18.2 Shannon's Entropy, 263

18.3 The Plug-in (or Maximum Likelihood) Estimator, 264

18.4 Lempel-Ziv Estimators, 265

18.5 Encoding Schemes, 269

18.5.1 Binary Encoding, 270

18.5.2 Quantile Encoding, 270

18.5.3 Sigma Encoding, 270

18.6 Entropy of a Gaussian Process, 271

18.7 Entropy and the Generalized Mean, 271

18.8 A Few Financial Applications of Entropy, 275

18.8.1 Market Efficiency, 275

18.8.2 Maximum Entropy Generation, 275

18.8.3 Portfolio Concentration, 275

18.8.4 Market Microstructure, 276

Exercises, 277

References, 278

Bibliography, 279

19 Microstructural Features 281

19.1 Motivation, 281

19.2 Review of the Literature, 281

19.3 First Generation: Price Sequences, 282

19.3.1 The Tick Rule, 282

19.3.2 The Roll Model, 282

19.3.3 High-Low Volatility Estimator, 283

19.3.4 Corwin and Schultz, 284

19.4 Second Generation: Strategic Trade Models, 286

19.4.1 Kyle's Lambda, 286

19.4.2 Amihud's Lambda, 288

19.4.3 Hasbrouck's Lambda, 289

19.5 Third Generation: Sequential Trade Models, 290

19.5.1 Probability of Information-based Trading, 290

19.5.2 Volume-Synchronized Probability of Informed Trading, 292

19.6 Additional Features from Microstructural Datasets, 293

19.6.1 Distibution of Order Sizes, 293

19.6.2 Cancellation Rates, Limit Orders, Market Orders, 293

19.6.3 Time-Weighted Average Price Execution Algorithms, 294

19.6.4 Options Markets, 295

19.6.5 Serial Correlation of Signed Order Flow, 295

19.7 What Is Microstructural Information?, 295

Exercises, 296

References, 298

PART 5 HIGH-PERFORMANCE COMPUTING RECIPES 301

20 Multiprocessing and Vectorization 303

20.1 Motivation, 303

20.2 Vectorization Example, 303

20.3 Single-Thread vs. Multithreading vs. Multiprocessing, 304

20.4 Atoms and Molecules, 306

20.4.1 Linear Partitions, 306

20.4.2 Two-Nested Loops Partitions, 307

20.5 Multiprocessing Engines, 309

20.5.1 Preparing the Jobs, 309

20.5.2 Asynchronous Calls, 311

20.5.3 Unwrapping the Callback, 312

20.5.4 Pickle/Unpickle Objects, 313

20.5.5 Output Reduction, 313

20.6 Multiprocessing Example, 315

Exercises, 316

Reference, 317
Bibliography, 317

21 Brute Force and Quantum Computers

21.1 Motivation, 319
21.2 Combinatorial Optimization, 319
21.3 The Objective Function, 320
21.4 The Problem, 321
21.5 An Integer Optimization Approach, 321
21.5.1 Pigeonhole Partitions, 321
21.5.2 Feasible Static Solutions, 323
21.5.3 Evaluating Trajectories, 323
21.6 A Numerical Example, 325
21.6.1 Random Matrices, 325
21.6.2 Static Solution, 326
21.6.3 Dynamic Solution, 327
Exercises, 327
References, 328

22 High-Performance Computational Intelligence and Forecasting Technologies
Kesheng Wu and Horst D. Simon

22.1 Motivation, 329
22.2 Regulatory Response to the Flash Crash of 2010, 329
22.3 Background, 330
22.4 HPC Hardware, 331
22.5 HPC Software, 335
22.5.1 Message Passing Interface, 335
22.5.2 Hierarchical Data Format 5, 336
22.5.3 *In Situ* Processing, 336
22.5.4 Convergence, 337
22.6 Use Cases, 337
22.6.1 Supernova Hunting, 337
22.6.2 Blobs in Fusion Plasma, 338
22.6.3 Intraday Peak Electricity Usage, 340
22.6.4 The Flash Crash of 2010, 341
22.6.5 Volume-synchronized Probability of Informed Trading Calibration, 346

22.6.6 Revealing High Frequency Events with Non-uniform Fast Fourier Transform, 347

22.7 Summary and Call for Participation, 349

22.8 Acknowledgments, 350

References, 350

Index **353**