

MARCOS LÓPEZ DE PRADO

ADVANCES
in
FINANCIAL
MACHINE
LEARNING

WILEY

Praise for *Advances in Financial Machine Learning*

In his new book *Advances in Financial Machine Learning*, noted financial scholar Marcos López de Prado strikes a well-aimed karate chop at the naive and often statistically overfit techniques that are so prevalent in the financial world today. He points out that not only are business-as-usual approaches largely impotent in today's high-tech finance, but in many cases they are actually prone to lose money. But López de Prado does more than just expose the mathematical and statistical sins of the finance world. Instead, he offers a technically sound roadmap for finance professionals to join the wave of machine learning. What is particularly refreshing is the author's empirical approach—his focus is on real-world data analysis, not on purely theoretical methods that may look pretty on paper but which, in many cases, are largely ineffective in practice. The book is geared to finance professionals who are already familiar with statistical data analysis techniques, but it is well worth the effort for those who want to do real state-of-the-art work in the field.”

Dr. David H. Bailey, former Complex Systems Lead,
Lawrence Berkeley National Laboratory. Co-discoverer of the
BBP spigot algorithm

“Finance has evolved from a compendium of heuristics based on historical financial statements to a highly sophisticated scientific discipline relying on computer farms to analyze massive data streams in real time. The recent highly impressive advances in machine learning (ML) are fraught with both promise and peril when applied to modern finance. While finance offers up the nonlinearities and large data sets upon which ML thrives, it also offers up noisy data and the human element which presently lie beyond the scope of standard ML techniques. To err is human, but if you really want to f**k things up, use a computer. Against this background, Dr. López de Prado has written the first comprehensive book describing the application of modern ML to financial modeling. The book blends the latest technological developments in ML with critical life lessons learned from the author's decades of financial experience in leading academic and industrial institutions. I highly recommend this exciting book to both prospective students of financial ML and the professors and supervisors who teach and guide them.”

Prof. Peter Carr, Chair of the Finance and Risk Engineering
Department, NYU Tandon School of Engineering

“Marcos is a visionary who works tirelessly to advance the finance field. His writing is comprehensive and masterfully connects the theory to the application. It is not often you find a book that can cross that divide. This book is an essential read for both practitioners and technologists working on solutions for the investment community.”

Landon Downs, President and Cofounder, 1QBit

“Academics who want to understand modern investment management need to read this book. In it, Marcos López de Prado explains how portfolio managers use machine learning to derive, test, and employ trading strategies. He does this from a very unusual combination of an academic perspective and extensive experience in industry, allowing him to both explain in detail what happens in industry and to explain

how it works. I suspect that some readers will find parts of the book that they do not understand or that they disagree with, but everyone interested in understanding the application of machine learning to finance will benefit from reading this book.”

Prof. David Easley, Cornell University. Chair of the
NASDAQ-OMX Economic Advisory Board

“For many decades, finance has relied on overly simplistic statistical techniques to identify patterns in data. Machine learning promises to change that by allowing researchers to use modern nonlinear and highly dimensional techniques, similar to those used in scientific fields like DNA analysis and astrophysics. At the same time, applying those machine learning algorithms to model financial problems would be dangerous. Financial problems require very distinct machine learning solutions. Dr. López de Prado’s book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book.”

Prof. Frank Fabozzi, EDHEC Business School. Editor of
The Journal of Portfolio Management

“This is a welcome departure from the knowledge hoarding that plagues quantitative finance. López de Prado defines for all readers the next era of finance: industrial scale scientific research powered by machines.”

John Fawcett, Founder and CEO, Quantopian

“Marcos has assembled in one place an invaluable set of lessons and techniques for practitioners seeking to deploy machine learning techniques in finance. If machine learning is a new and potentially powerful weapon in the arsenal of quantitative finance, Marcos’s insightful book is laden with useful advice to help keep a curious practitioner from going down any number of blind alleys, or shooting oneself in the foot.”

Ross Garon, Head of Cubist Systematic Strategies. Managing
Director, Point72 Asset Management

“The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave, and it will touch every aspect of finance. López de Prado’s *Advances in Financial Machine Learning* is essential for readers who want to be ahead of the technology rather than being replaced by it.”

Prof. Campbell Harvey, Duke University. Former President of
the American Finance Association

“How does one make sense of today’s financial markets in which complex algorithms route orders, financial data is voluminous, and trading speeds are measured in nanoseconds? In this important book, Marcos López de Prado sets out a new paradigm for investment management built on machine learning. Far from being a “black box” technique, this book clearly explains the tools and process of financial

machine learning. For academics and practitioners alike, this book fills an important gap in our understanding of investment management in the machine age.”

Prof. Maureen O’Hara, Cornell University. Former President of the American Finance Association

“Marcos López de Prado has produced an extremely timely and important book on machine learning. The author’s academic and professional first-rate credentials shine through the pages of this book—indeed, I could think of few, if any, authors better suited to explaining both the theoretical and the practical aspects of this new and (for most) unfamiliar subject. Both novices and experienced professionals will find insightful ideas, and will understand how the subject can be applied in novel and useful ways. The Python code will give the novice readers a running start and will allow them to gain quickly a hands-on appreciation of the subject. Destined to become a classic in this rapidly burgeoning field.”

Prof. Riccardo Rebonato, EDHEC Business School. Former Global Head of Rates and FX Analytics at PIMCO

“A tour de force on practical aspects of machine learning in finance, brimming with ideas on how to employ cutting-edge techniques, such as fractional differentiation and quantum computers, to gain insight and competitive advantage. A useful volume for finance and machine learning practitioners alike.”

Dr. Collin P. Williams, Head of Research, D-Wave Systems

(c) 2018 by Marcos Lopez de Prado. Reprinted with permission. All rights reserved. Full version available at <https://goo.gl/w6gMdq>

(c) 2018 by Marcos Lopez de Prado. Reprinted with permission. All rights reserved. Full version available at <https://goo.gl/w6gMdq>

Advances in Financial Machine Learning

(c) 2018 by Marcos Lopez de Prado. Reprinted with permission. All rights reserved. Full version available at <https://goo.gl/w6gMdq>

(c) 2018 by Marcos Lopez de Prado. Reprinted with permission. All rights reserved. Full version available at <https://goo.gl/w6gMdq>

Advances in Financial Machine Learning

MARCOS LÓPEZ DE PRADO

WILEY

Cover image: © Erikona/Getty Images

Cover design: Wiley

Copyright © 2018 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

The views expressed in this book are the author's and do not necessarily reflect those of the organizations he is affiliated with.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the Web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at www.wiley.com/go/permissions.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. The views expressed in this book are the author's and do not necessarily reflect those of the organizations he is affiliated with.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993, or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

ISBN 978-1-119-48208-6 (Hardcover)

ISBN 978-1-119-48211-6 (ePDF)

ISBN 978-1-119-48210-9 (ePub)

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Dedicated to the memory of my coauthor and friend,
Professor Jonathan M. Borwein, FRSC, FAAAS,
FBAS, FAustMS, FAA, FAMS, FRSNSW
(1951–2016)

(c) 2018 by Marcos Lopez de Prado. Reprinted with permission. All rights reserved. Full version available at <https://goo.gl/w6gMdq>

There are very few things which we know, which are not capable of being reduced to a mathematical reasoning. And when they cannot, it's a sign our knowledge of them is very small and confused. Where a mathematical reasoning can be had, it's as great a folly to make use of any other, as to grope for a thing in the dark, when you have a candle standing by you.

—*Of the Laws of Chance*, Preface (1692)
John Arbuthnot (1667–1735)

(c) 2018 by Marcos Lopez de Prado. Reprinted with permission. All rights reserved. Full version available at <https://goo.gl/w6gMdq>

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

3 Labeling 43

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

4 Sample Weights 59

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	103
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	113
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 Tuning with Cross-Validation	129
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	169
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	195
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 Implementable, 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	Distribution 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 **319**

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 **329**

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

CONTENTS	xix
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

(c) 2018 by Marcos Lopez de Prado. Reprinted with permission. All rights reserved. Full version available at <https://goo.gl/w6gMdq>

About the Author

Marcos López de Prado manages several multibillion-dollar funds for institutional investors using machine learning algorithms. Over the past 20 years, his work has combined advanced mathematics with supercomputing technologies to deliver billions of dollars in net profits for investors and firms. A proponent of research by collaboration, Marcos has published with more than 30 leading academics, resulting in some of the most-read papers in finance.

Since 2010, Marcos has also been a Research Fellow at Lawrence Berkeley National Laboratory (U.S. Department of Energy's Office of Science), where he conducts research focused on the mathematics of large-scale financial problems and high-performance computing at the Computational Research department. For the past seven years he has lectured at Cornell University, where he currently teaches a graduate course in financial big data and machine learning in the Operations Research department.

Marcos is the recipient of the 1999 National Award for Academic Excellence, which the government of Spain bestows upon the best graduate student nationally. He earned a PhD in financial economics (2003) and a second PhD in mathematical finance (2011) from Universidad Complutense de Madrid. Between his two doctorates, Marcos was a postdoctoral research fellow of RCC at Harvard University for three years, during which he published more than a dozen articles in JCR-indexed scientific journals. Marcos has an Erdős #2 and an Einstein #4, according to the American Mathematical Society.

(c) 2018 by Marcos Lopez de Prado. Reprinted with permission. All rights reserved. Full version available at <https://goo.gl/w6gMdq>

CHAPTER 1

Financial Machine Learning as a Distinct Subject

1.1 MOTIVATION

Machine learning (ML) is changing virtually every aspect of our lives. Today ML algorithms accomplish tasks that until recently only expert humans could perform. As it relates to finance, this is the most exciting time to adopt a disruptive technology that will transform how everyone invests for generations. This book explains scientifically sound ML tools that have worked for me over the course of two decades, and have helped me to manage large pools of funds for some of the most demanding institutional investors.

Books about investments largely fall in one of two categories. On one hand we find books written by authors who have not practiced what they teach. They contain extremely elegant mathematics that describes a world that does not exist. Just because a theorem is true in a logical sense does not mean it is true in a physical sense. On the other hand we find books written by authors who offer explanations absent of any rigorous academic theory. They misuse mathematical tools to describe actual observations. Their models are overfit and fail when implemented. Academic investigation and publication are divorced from practical application to financial markets, and many applications in the trading/investment world are not grounded in proper science.

A first motivation for writing this book is to cross the proverbial divide that separates academia and the industry. I have been on both sides of the rift, and I understand how difficult it is to cross it and how easy it is to get entrenched on one side. Virtue is in the balance. This book will not advocate a theory merely because of its mathematical beauty, and will not propose a solution just because it appears to work. My goal is to transmit the kind of knowledge that only comes from experience, formalized in a rigorous manner.

A second motivation is inspired by the desire that finance serves a purpose. Over the years some of my articles, published in academic journals and newspapers, have expressed my displeasure with the current role that finance plays in our society. Investors are lured to gamble their wealth on wild hunches originated by charlatans and encouraged by mass media. One day in the near future, ML will dominate finance, science will curtail guessing, and investing will not mean gambling. I would like the reader to play a part in that revolution.

A third motivation is that many investors fail to grasp the complexity of ML applications to investments. This seems to be particularly true for discretionary firms moving into the “quantamental” space. I am afraid their high expectations will not be met, not because ML failed, but because they used ML incorrectly. Over the coming years, many firms will invest with off-the-shelf ML algorithms, directly imported from academia or Silicon Valley, and my forecast is that they will lose money (to better ML solutions). Beating the wisdom of the crowds is harder than recognizing faces or driving cars. With this book my hope is that you will learn how to solve some of the challenges that make finance a particularly difficult playground for ML, like backtest overfitting. Financial ML is a subject in its own right, related to but separate from standard ML, and this book unravels it for you.

1.2 THE MAIN REASON FINANCIAL MACHINE LEARNING PROJECTS USUALLY FAIL

The rate of failure in quantitative finance is high, particularly so in financial ML. The few who succeed amass a large amount of assets and deliver consistently exceptional performance to their investors. However, that is a rare outcome, for reasons explained in this book. Over the past two decades, I have seen many faces come and go, firms started and shut down. In my experience, there is one critical mistake that underlies all those failures.

1.2.1 The Sisyphus Paradigm

Discretionary portfolio managers (PMs) make investment decisions that do not follow a particular theory or rationale (if there were one, they would be systematic PMs). They consume raw news and analyses, but mostly rely on their judgment or intuition. They may rationalize those decisions based on some story, but there is always a story for every decision. Because nobody fully understands the logic behind their bets, investment firms ask them to work independently from one another, in silos, to ensure diversification. If you have ever attended a meeting of discretionary PMs, you probably noticed how long and aimless they can be. Each attendee seems obsessed about one particular piece of anecdotal information, and giant argumentative leaps are made without fact-based, empirical evidence. This does not mean that discretionary PMs cannot be successful. On the contrary, a few of them are. The point is, they cannot naturally work as a team. Bring 50 discretionary PMs together, and they

will influence one another until eventually you are paying 50 salaries for the work of one. Thus it makes sense for them to work in silos so they interact as little as possible.

Wherever I have seen that formula applied to quantitative or ML projects, it has led to disaster. The boardroom's mentality is, let us do with quants what has worked with discretionary PMs. Let us hire 50 PhDs and demand that each of them produce an investment strategy within six months. This approach always backfires, because each PhD will frantically search for investment opportunities and eventually settle for (1) a false positive that looks great in an overfit backtest or (2) standard factor investing, which is an overcrowded strategy with a low Sharpe ratio, but at least has academic support. Both outcomes will disappoint the investment board, and the project will be cancelled. Even if 5 of those PhDs identified a true discovery, the profits would not suffice to cover for the expenses of 50, so those 5 will relocate somewhere else, searching for a proper reward.

1.2.2 The Meta-Strategy Paradigm

If you have been asked to develop ML strategies on your own, the odds are stacked against you. It takes almost as much effort to produce one true investment strategy as to produce a hundred, and the complexities are overwhelming: data curation and processing, HPC infrastructure, software development, feature analysis, execution simulators, backtesting, etc. Even if the firm provides you with shared services in those areas, you are like a worker at a BMW factory who has been asked to build an entire car by using all the workshops around you. One week you need to be a master welder, another week an electrician, another week a mechanical engineer, another week a painter . . . You will try, fail, and circle back to welding. How does that make sense?

Every successful quantitative firm I am aware of applies the meta-strategy paradigm (López de Prado [2014]). Accordingly, this book was written as a research manual for teams, not for individuals. Through its chapters you will learn how to set up a research factory, as well as the various stations of the assembly line. The role of each quant is to specialize in a particular task, to become the best there is at it, while having a holistic view of the entire process. This book outlines the factory plan, where teamwork yields discoveries at a predictable rate, with no reliance on lucky strikes. This is how Berkeley Lab and other U.S. National Laboratories routinely make scientific discoveries, such as adding 16 elements to the periodic table, or laying out the groundwork for MRIs and PET scans.¹ No particular individual is responsible for these discoveries, as they are the outcome of team efforts where everyone contributes. Of course, setting up these financial laboratories takes time, and requires people who know what they are doing and have done it before. But what do you think has a higher chance of success, this proven paradigm of organized collaboration or the Sisyphean alternative of having every single quant rolling their immense boulder up the mountain?

¹ Berkeley Lab, <http://www.lbl.gov/about>.

1.3 BOOK STRUCTURE

This book disentangles a web of interconnected topics and presents them in an ordered fashion. Each chapter assumes that you have read the previous ones. Part 1 will help you structure your financial data in a way that is amenable to ML algorithms. Part 2 discusses how to do research with ML algorithms on that data. Here the emphasis is on doing research and making an actual discovery through a scientific process, as opposed to searching aimlessly until some serendipitous (likely false) result pops up. Part 3 explains how to backtest your discovery and evaluate the probability that it is false.

These three parts give an overview of the entire process, from data analysis to model research to discovery evaluation. With that knowledge, Part 4 goes back to the data and explains innovative ways to extract informative features. Finally, much of this work requires a lot of computational power, so Part 5 wraps up the book with some useful HPC recipes.

1.3.1 Structure by Production Chain

Mining gold or silver was a relatively straightforward endeavor during the 16th and 17th centuries. In less than a hundred years, the Spanish treasure fleet quadrupled the amount of precious metals in circulation throughout Europe. Those times are long gone, and today prospectors must deploy complex industrial methods to extract microscopic bullion particles out of tons of earth. That does not mean that gold production is at historical lows. On the contrary, nowadays miners extract 2,500 metric tons of microscopic gold every year, compared to the average annual 1.54 metric tons taken by the Spanish conquistadors throughout the entire 16th century!² Visible gold is an infinitesimal portion of the overall amount of gold on Earth. *El Dorado* was always there . . . if only Pizarro could have exchanged the sword for a microscope.

The discovery of investment strategies has undergone a similar evolution. If a decade ago it was relatively common for an individual to discover macroscopic alpha (i.e., using simple mathematical tools like econometrics), currently the chances of that happening are quickly converging to zero. Individuals searching nowadays for macroscopic alpha, regardless of their experience or knowledge, are fighting overwhelming odds. The only true alpha left is microscopic, and finding it requires capital-intensive industrial methods. Just like with gold, microscopic alpha does not mean smaller overall profits. Microscopic alpha today is much more abundant than macroscopic alpha has ever been in history. There is a lot of money to be made, but you will need to use heavy ML tools.

Let us review some of the stations involved in the chain of production within a modern asset manager.

² <http://www.numbersleuth.org/worlds-gold/>.

1.3.1.1 Data Curators

This is the station responsible for collecting, cleaning, indexing, storing, adjusting, and delivering all data to the production chain. The values could be tabulated or hierarchical, aligned or misaligned, historical or real-time feeds, etc. Team members are experts in market microstructure and data protocols such as FIX. They must develop the data handlers needed to understand the context in which that data arises. For example, was a quote cancelled and replaced at a different level, or cancelled without replacement? Each asset class has its own nuances. For instance, bonds are routinely exchanged or recalled; stocks are subjected to splits, reverse-splits, voting rights, etc.; futures and options must be rolled; currencies are not traded in a centralized order book. The degree of specialization involved in this station is beyond the scope of this book, and Chapter 1 will discuss only a few aspects of data curation.

1.3.1.2 Feature Analysts

This is the station responsible for transforming raw data into informative signals. These informative signals have some predictive power over financial variables. Team members are experts in information theory, signal extraction and processing, visualization, labeling, weighting, classifiers, and feature importance techniques. For example, feature analysts may discover that the probability of a sell-off is particularly high when: (1) quoted offers are cancelled-replaced with market sell orders, and (2) quoted buy orders are cancelled-replaced with limit buy orders deeper in the book. Such a finding is not an investment strategy on its own, and can be used in alternative ways: execution, monitoring of liquidity risk, market making, position taking, etc. A common error is to believe that feature analysts develop strategies. Instead, feature analysts collect and catalogue libraries of findings that can be useful to a multiplicity of stations. Chapters 2–9 and 17–19 are dedicated to this all-important station.

1.3.1.3 Strategists

In this station, informative features are transformed into actual investment algorithms. A strategist will parse through the libraries of features looking for ideas to develop an investment strategy. These features were discovered by different analysts studying a wide range of instruments and asset classes. The goal of the strategist is to make sense of all these observations and to formulate a general theory that explains them. Therefore, the strategy is merely the experiment designed to test the validity of this theory. Team members are data scientists with a deep knowledge of financial markets and the economy. Remember, the theory needs to explain a large collection of important features. In particular, a theory must identify the economic mechanism that causes an agent to lose money to us. Is it a behavioral bias? Asymmetric information? Regulatory constraints? Features may be discovered by a black box, but the strategy is developed in a white box. Gluing together a number of catalogued features does not constitute a theory. Once a strategy is finalized, the strategists will prepare code that utilizes the full algorithm and submit that prototype to the backtesting team described below. Chapters 10 and 16 are dedicated to this station, with the understanding that it would be unreasonable for a book to reveal specific investment strategies.

1.3.1.4 *Backtesters*

This station assesses the profitability of an investment strategy under various scenarios. One of the scenarios of interest is how the strategy would perform if history repeated itself. However, the historical path is merely one of the possible outcomes of a stochastic process, and not necessarily the most likely going forward. Alternative scenarios must be evaluated, consistent with the knowledge of the weaknesses and strengths of a proposed strategy. Team members are data scientists with a deep understanding of empirical and experimental techniques. A good backtester incorporates in his analysis meta-information regarding how the strategy came about. In particular, his analysis must evaluate the probability of backtest overfitting by taking into account the number of trials it took to distill the strategy. The results of this evaluation will not be reused by other stations, for reasons that will become apparent in Chapter 11. Instead, backtest results are communicated to management and not shared with anyone else. Chapters 11–16 discuss the analyses carried out by this station.

1.3.1.5 *Deployment Team*

The deployment team is tasked with integrating the strategy code into the production line. Some components may be reused by multiple strategies, especially when they share common features. Team members are algorithm specialists and hardcore mathematical programmers. Part of their job is to ensure that the deployed solution is logically identical to the prototype they received. It is also the deployment team's responsibility to optimize the implementation sufficiently, such that production latency is minimized. As production calculations often are time sensitive, this team will rely heavily on process schedulers, automation servers (Jenkins), vectorization, multithreading, multiprocessing, graphics processing unit (GPU-NVIDIA), distributed computing (Hadoop), high-performance computing (Slurm), and parallel computing techniques in general. Chapters 20–22 touch on various aspects interesting to this station, as they relate to financial ML.

1.3.1.6 *Portfolio Oversight*

Once a strategy is deployed, it follows a *cursus honorum*, which entails the following stages or lifecycle:

1. **Embargo:** Initially, the strategy is run on data observed after the end date of the backtest. Such a period may have been reserved by the backtesters, or it may be the result of implementation delays. If embargoed performance is consistent with backtest results, the strategy is promoted to the next stage.
2. **Paper trading:** At this point, the strategy is run on a live, real-time feed. In this way, performance will account for data parsing latencies, calculation latencies, execution delays, and other time lapses between observation and positioning. Paper trading will take place for as long as it is needed to gather enough evidence that the strategy performs as expected.
3. **Graduation:** At this stage, the strategy manages a real position, whether in isolation or as part of an ensemble. Performance is evaluated precisely, including attributed risk, returns, and costs.

4. **Re-allocation:** Based on the production performance, the allocation to graduated strategies is re-assessed frequently and automatically in the context of a diversified portfolio. In general, a strategy's allocation follows a concave function. The initial allocation (at graduation) is small. As time passes, and the strategy performs as expected, the allocation is increased. Over time, performance decays, and allocations become gradually smaller.
5. **Decommission:** Eventually, all strategies are discontinued. This happens when they perform below expectations for a sufficiently extended period of time to conclude that the supporting theory is no longer backed by empirical evidence.

In general, it is preferable to release new variations of a strategy and run them in parallel with old versions. Each version will go through the above lifecycle, and old strategies will receive smaller allocations as a matter of diversification, while taking into account the degree of confidence derived from their longer track record.

1.3.2 Structure by Strategy Component

Many investment managers believe that the secret to riches is to implement an extremely complex ML algorithm. They are setting themselves up for a disappointment. If it was as easy as coding a state-of-the-art classifier, most people in Silicon Valley would be billionaires. A successful investment strategy is the result of multiple factors. Table 1.1 summarizes what chapters will help you address each of the challenges involved in developing a successful investment strategy.

Throughout the book, you will find many references to journal articles I have published over the years. Rather than repeating myself, I will often refer you to one of them, where you will find a detailed analysis of the subject at hand. All of my cited papers can be downloaded for free, in pre-print format, from my website: www.QuantResearch.org.

1.3.2.1 Data

- Problem: Garbage in, garbage out.
- Solution: Work with unique, hard-to-manipulate data. If you are the only user of this data, whatever its value, it is all for you.
- How:
 - Chapter 2: Structure your data correctly.
 - Chapter 3: Produce informative labels.
 - Chapters 4 and 5: Model non-IID series properly.
 - Chapters 17–19: Find predictive features.

1.3.2.2 Software

- Problem: A specialized task requires customized tools.
- Solution: Develop your own classes. Using popular libraries means more competitors tapping the same well.

TABLE 1.1 Overview of the Challenges Addressed by Every Chapter

Part	Chapter	Fin. data	Software	Hardware	Math	Meta-Strat	Overfitting
1	2	X	X				
1	3	X	X				
1	4	X	X				
1	5	X	X		X		
2	6		X				
2	7		X			X	X
2	8		X			X	
2	9		X			X	
3	10		X			X	
3	11		X		X		X
3	12		X		X		X
3	13		X		X		X
3	14		X		X		X
3	15		X		X		X
3	16		X		X	X	X
4	17	X	X		X		
4	18	X	X		X		
4	19	X	X				
5	20		X	X	X		
5	21		X	X	X		
5	22		X	X	X		

- How:
 - Chapters 2–22: Throughout the book, for each chapter, we develop our own functions. For your particular problems, you will have to do the same, following the examples in the book.

1.3.2.3 Hardware

- Problem: ML involves some of the most computationally intensive tasks in all of mathematics.
- Solution: Become an HPC expert. If possible, partner with a National Laboratory to build a supercomputer.
- How:
 - Chapters 20 and 22: Learn how to think in terms of multiprocessing architectures. Whenever you code a library, structure it in such a way that functions can be called in parallel. You will find plenty of examples in the book.
 - Chapter 21: Develop algorithms for quantum computers.

1.3.2.4 Math

- Problem: Mathematical proofs can take years, decades, and centuries. No investor will wait that long.

- Solution: Use experimental math. Solve hard, intractable problems, not by proof but by experiment. For example, Bailey, Borwein, and Plouffe [1997] found a spigot algorithm for π (pi) without proof, against the prior perception that such mathematical finding would not be possible.
- How:
 - Chapter 5: Familiarize yourself with memory-preserving data transformations.
 - Chapters 11–15: There are experimental methods to assess the value of your strategy, with greater reliability than a historical simulation.
 - Chapter 16: An algorithm that is optimal in-sample can perform poorly out-of-sample. There is no mathematical proof for investment success. Rely on experimental methods to lead your research.
 - Chapters 17 and 18: Apply methods to detect structural breaks, and quantify the amount of information carried by financial series.
 - Chapter 20: Learn queuing methods for distributed computing so that you can break apart complex tasks and speed up calculations.
 - Chapter 21: Become familiar with discrete methods, used among others by quantum computers, to solve intractable problems.

1.3.2.5 *Meta-Strategies*

- Problem: Amateurs develop individual strategies, believing that there is such a thing as a magical formula for riches. In contrast, professionals develop methods to mass-produce strategies. The money is not in making a car, it is in making a car factory.
- Solution: Think like a business. Your goal is to run a research lab like a factory, where true discoveries are not born out of inspiration, but out of methodic hard work. That was the philosophy of physicist Ernest Lawrence, the founder of the first U.S. National Laboratory.
- How:
 - Chapters 7–9: Build a research process that identifies features relevant across asset classes, while dealing with multi-collinearity of financial features.
 - Chapter 10: Combine multiple predictions into a single bet.
 - Chapter 16: Allocate funds to strategies using a robust method that performs well out-of-sample.

1.3.2.6 *Overfitting*

- Problem: Standard cross-validation methods fail in finance. Most discoveries in finance are false, due to multiple testing and selection bias.
- Solution:
 - Whatever you do, always ask yourself in what way you may be overfitting. Be skeptical about your own work, and constantly challenge yourself to prove that you are adding value.

- Overfitting is unethical. It leads to promising outcomes that cannot be delivered. When done knowingly, overfitting is outright scientific fraud. The fact that many academics do it does not make it right: They are not risking anyone's wealth, not even theirs.
- It is also a waste of your time, resources, and opportunities. Besides, the industry only pays for out-of-sample returns. You will only succeed *after* you have created substantial wealth for your investors.
- How:
 - Chapters 11–15: There are three backtesting paradigms, of which historical simulation is only one. Each backtest is always overfit to some extent, and it is critical to learn to quantify by how much.
 - Chapter 16: Learn robust techniques for asset allocation that do not overfit in-sample signals at the expense of out-of-sample performance.

1.3.3 Structure by Common Pitfall

Despite its many advantages, ML is no panacea. The flexibility and power of ML techniques have a dark side. When misused, ML algorithms will confuse statistical flukes with patterns. This fact, combined with the low signal-to-noise ratio that characterizes finance, all but ensures that careless users will produce false discoveries at an ever-greater speed. This book exposes some of the most pervasive errors made by ML experts when they apply their techniques on financial datasets. Some of these pitfalls are listed in Table 1.2, with solutions that are explained in the indicated chapters.

1.4 TARGET AUDIENCE

This book presents advanced ML methods specifically designed to address the challenges posed by financial datasets. By “advanced” I do not mean extremely difficult to grasp, or explaining the latest reincarnation of deep, recurrent, or convolutional neural networks. Instead, the book answers questions that senior researchers, who have experience applying ML algorithms to financial problems, will recognize as critical. If you are new to ML, and you do not have experience working with complex algorithms, this book may not be for you (yet). Unless you have confronted in practice the problems discussed in these chapters, you may have difficulty understanding the utility of solving them. Before reading this book, you may want to study several excellent introductory ML books published in recent years. I have listed a few of them in the references section.

The core audience of this book is investment professionals with a strong ML background. My goals are that you monetize what you learn in this book, help us modernize finance, and deliver actual value for investors.

This book also targets data scientists who have successfully implemented ML algorithms in a variety of fields outside finance. If you have worked at Google and have applied deep neural networks to face recognition, but things do not seem to

TABLE 1.2 Common Pitfalls in Financial ML

#	Category	Pitfall	Solution	Chapter
1	Epistemological	The Sisyphus paradigm	The meta-strategy paradigm	1
2	Epistemological	Research through backtesting	Feature importance analysis	8
3	Data processing	Chronological sampling	The volume clock	2
4	Data processing	Integer differentiation	Fractional differentiation	5
5	Classification	Fixed-time horizon labeling	The triple-barrier method	3
6	Classification	Learning side and size simultaneously	Meta-labeling	3
7	Classification	Weighting of non-IID samples	Uniqueness weighting; sequential bootstrapping	4
8	Evaluation	Cross-validation leakage	Purging and embargoing	7,9
9	Evaluation	Walk-forward (historical) backtesting	Combinatorial purged cross-validation	11,12
10	Evaluation	Backtest overfitting	Backtesting on synthetic data; the deflated Sharpe ratio	10–16

work so well when you run your algorithms on financial data, this book will help you. Sometimes you may not understand the financial rationale behind some structures (e.g., meta-labeling, the triple-barrier method, fracdiff), but bear with me: Once you have managed an investment portfolio long enough, the rules of the game will become clearer to you, along with the meaning of these chapters.

1.5 REQUISITES

Investment management is one of the most multi-disciplinary areas of research, and this book reflects that fact. Understanding the various sections requires a practical knowledge of ML, market microstructure, portfolio management, mathematical finance, statistics, econometrics, linear algebra, convex optimization, discrete math, signal processing, information theory, object-oriented programming, parallel processing, and supercomputing.

Python has become the *de facto* standard language for ML, and I have to assume that you are an experienced developer. You must be familiar with scikit-learn (sklearn), pandas, numpy, scipy, multiprocessing, matplotlib and a few other libraries.

Code snippets invoke functions from these libraries using their conventional prefix, `pd` for `pandas`, `np` for `numpy`, `mpl` for `matplotlib`, etc. There are numerous books on each of these libraries, and you cannot know enough about the specifics of each one. Throughout the book we will discuss some issues with their implementation, including unresolved bugs to keep in mind.

1.6 FAQs

How can ML algorithms be useful in finance?

Many financial operations require making decisions based on pre-defined rules, like option pricing, algorithmic execution, or risk monitoring. This is where the bulk of automation has taken place so far, transforming the financial markets into ultra-fast, hyper-connected networks for exchanging information. In performing these tasks, machines were asked to follow the rules as fast as possible. High-frequency trading is a prime example. See Easley, López de Prado, and O'Hara [2013] for a detailed treatment of the subject.

The algorithmization of finance is unstoppable. Between June 12, 1968, and December 31, 1968, the NYSE was closed every Wednesday, so that back office could catch up with paperwork. Can you imagine that? We live in a different world today, and in 10 years things will be even better. Because the next wave of automation does not involve following rules, but making judgment calls. As emotional beings, subject to fears, hopes, and agendas, humans are not particularly good at making fact-based decisions, particularly when those decisions involve conflicts of interest. In those situations, investors are better served when a machine makes the calls, based on facts learned from hard data. This not only applies to investment strategy development, but to virtually every area of financial advice: granting a loan, rating a bond, classifying a company, recruiting talent, predicting earnings, forecasting inflation, etc. Furthermore, machines will comply with the law, always, when programmed to do so. If a dubious decision is made, investors can go back to the logs and understand exactly what happened. It is much easier to improve an algorithmic investment process than one relying entirely on humans.

How can ML algorithms beat humans at investing?

Do you remember when people were certain that computers would never beat humans at chess? Or *Jeopardy!*? Poker? Go? Millions of years of evolution (a genetic algorithm) have fine-tuned our ape brains to survive in a hostile 3-dimensional world where the laws of nature are static. Now, when it comes to identifying subtle patterns in a high-dimensional world, where the rules of the game change every day, all that fine-tuning turns out to be detrimental. An ML algorithm can spot patterns in a 100-dimensional world as easily as in our familiar 3-dimensional one. And while we all laugh when we see an algorithm make a silly mistake, keep in mind, algorithms have been around only a fraction of our millions of years. Every day they get better at this, we do not. Humans are slow learners, which puts us at a disadvantage in a fast-changing world like finance.

Does that mean that there is no space left for human investors?

Not at all. No human is better at chess than a computer. And no computer is better at chess than a human supported by a computer. Discretionary PMs are at a disadvantage when betting against an ML algorithm, but it is possible that the best results are achieved by combining discretionary PMs with ML algorithms. This is what has come to be known as the “quantamental” way. Throughout the book you will find techniques that can be used by quantamental teams, that is, methods that allow you to combine human guesses (inspired by fundamental variables) with mathematical forecasts. In particular, Chapter 3 introduces a new technique called meta-labeling, which allows you to add an ML layer on top of a discretionary one.

How does financial ML differ from econometrics?

Econometrics is the application of classical statistical methods to economic and financial series. The essential tool of econometrics is multivariate linear regression, an 18th-century technology that was already mastered by Gauss before 1794 (Stigler [1981]). Standard econometric models do not learn. It is hard to believe that something as complex as 21st-century finance could be grasped by something as simple as inverting a covariance matrix.

Every empirical science must build theories based on observation. If the statistical toolbox used to model these observations is linear regression, the researcher will fail to recognize the complexity of the data, and the theories will be awfully simplistic, useless. I have no doubt in my mind, econometrics is a primary reason economics and finance have not experienced meaningful progress over the past 70 years (Calkin and López de Prado [2014a, 2014b]).

For centuries, medieval astronomers made observations and developed theories about celestial mechanics. These theories never considered non-circular orbits, because they were deemed unholy and beneath God’s plan. The prediction errors were so gross, that ever more complex theories had to be devised to account for them. It was not until Kepler had the temerity to consider non-circular (elliptical) orbits that all of the sudden a much simpler general model was able to predict the position of the planets with astonishing accuracy. What if astronomers had never considered non-circular orbits? Well . . . what if economists finally started to consider non-linear functions? Where is our Kepler? Finance does not have a *Principia* because no Kepler means no Newton.

Financial ML methods do not replace theory. They guide it. An ML algorithm learns patterns in a high-dimensional space without being specifically directed. Once we understand what features are predictive of a phenomenon, we can build a theoretical explanation, which can be tested on an independent dataset. Students of economics and finance would do well enrolling in ML courses, rather than econometrics. Econometrics may be good enough to succeed in financial academia (for now), but succeeding in business requires ML.

What do you say to people who dismiss ML algorithms as black boxes?

If you are reading this book, chances are ML algorithms are white boxes to you. They are transparent, well-defined, crystal-clear, pattern-recognition functions. Most

people do not have your knowledge, and to them ML is like a magician's box: "Where did that rabbit come from? How are you tricking us, witch?" People mistrust what they do not understand. Their prejudices are rooted in ignorance, for which the Socratic remedy is simple: education. Besides, some of us enjoy using our brains, even though neuroscientists still have not figured out exactly how they work (a black box in itself).

From time to time you will encounter Luddites, who are beyond redemption. Ned Ludd was a weaver from Leicester, England, who in 1779 smashed two knitting frames in an outrage. With the advent of the industrial revolution, mobs infuriated by mechanization sabotaged and destroyed all machinery they could find. Textile workers ruined so much industrial equipment that Parliament had to pass laws making "machine breaking" a capital crime. Between 1811 and 1816, large parts of England were in open rebellion, to the point that there were more British troops fighting Luddites than there were fighting Napoleon on the Iberian Peninsula. The Luddite rebellion ended with brutal suppression through military force. Let us hope that the black box movement does not come to that.

Why don't you discuss specific ML algorithms?

The book is agnostic with regards to the particular ML algorithm you choose. Whether you use convolutional neural networks, AdaBoost, RFs, SVMs, and so on, there are many shared generic problems you will face: data structuring, labeling, weighting, stationary transformations, cross-validation, feature selection, feature importance, overfitting, backtesting, etc. In the context of financial modeling, answering these questions is non-trivial, and framework-specific approaches need to be developed. That is the focus of this book.

What other books do you recommend on this subject?

To my knowledge, this is the first book to provide a complete and systematic treatment of ML methods specific for finance: starting with a chapter dedicated to financial data structures, another chapter for labeling of financial series, another for sample weighting, time series differentiation, . . . all the way to a full part devoted to the proper backtesting of investment strategies. To be sure, there are a handful of prior publications (mostly journal articles) that have applied standard ML to financial series, but that is not what this book offers. My goal has been to address the unique nuisances that make financial ML modeling particularly challenging. Like any new subject, it is fast evolving, and the book will be updated as major advances take place. Please contact me at mldp@quantresearch.org if there is any particular topic you would like to see treated in future editions. I will gladly add those chapters, while acknowledging the names of those readers who suggested them.

I do not understand some of the sections and chapters. What should I do?

My advice is that you start by reading the references listed at the end of the chapter. When I wrote the book, I had to assume the reader was familiar with the existing literature, or this book would lose its focus. If after reading those references the sections still do not make sense, the likely reason is that they are related to a problem well understood by investment professionals (even if there is no mention of it in the

literature). For example, Chapter 2 will discuss effective methods to adjust futures prices for the roll, a problem known to most practitioners, even though it is rarely addressed in textbooks. I would encourage you to attend one of my regular seminars, and ask me your question at the end of my talk.

Why is the book so fixated on backtest overfitting?

There are two reasons. First, backtest overfitting is arguably the most important open problem in all of mathematical finance. It is our equivalent to “P versus NP” in computer science. If there was a precise method to prevent backtest overfitting, we would be able to take backtests to the bank. A backtest would be almost as good as cash, rather than a sales pitch. Hedge funds would allocate funds to portfolio managers with confidence. Investors would risk less, and would be willing to pay higher fees. Regulators would grant licenses to hedge fund managers on the basis of reliable evidence of skill and knowledge, leaving no space for charlatans. In my opinion, an investments book that does not address this issue is not worth your time. Why would you read a book that deals with CAPM, APT, asset allocation techniques, risk management, etc. when the empirical results that support those arguments were selected without determining their false discovery probabilities?

The second reason is that ML is a great weapon in your research arsenal, and a dangerous one to be sure. If backtest overfitting is an issue in econometric analysis, the flexibility of ML makes it a constant threat to your work. This is particularly the case in finance, because our datasets are shorter, with lower signal-to-noise ratio, and we do not have laboratories where we can conduct experiments while controlling for all environmental variables (López de Prado [2015]). An ML book that does not tackle these concerns can be more detrimental than beneficial to your career.

What is the mathematical nomenclature of the book?

When I started to write this book, I thought about assigning one symbol to each mathematical variable or function through all the chapters. That would work well if this book dealt with a single subject, like stochastic optimal control. However this book deals with a wide range of mathematical subjects, each with its own conventions. Readers would find it harder to consult references unless I also followed literature standards, which means that sometimes we must re-use symbols. To prevent any confusion, every chapter explains the nomenclature as it is being used. Most of the math is accompanied by a code snippet, so in case of doubt, please always follow the code.

Who wrote Chapter 22?

A popular perception is that ML is a new fascinating technology invented or perfected at IBM, Google, Facebook, Amazon, Netflix, Tesla, etc. It is true that technology firms have become heavy users of ML, especially in recent years. Those firms sponsored some of the most publicized recent ML achievements (like *Jeopardy!* or Go), which may have reinforced that perception.

However, the reader may be surprised to learn that, in fact, U.S. National Laboratories are among the research centers with the longest track record and experience

in using ML. These centers utilized ML before it was cool, and they applied it successfully for many decades to produce astounding scientific discoveries. If predicting what movies Netflix should recommend you to watch next is a worthy endeavor, so it is to understand the rate of expansion of the universe, or forecasting what coastlines will be most impacted by global warming, or preventing a cataclysmic failure of our national power grid. These are just some of the amazing questions that institutions like Berkeley Lab work on every day, quietly but tirelessly, with the help of ML.

In Chapter 22, Drs. Horst Simon and Kesheng Wu offer the perspective of a deputy director and a project leader at a major U.S. National Laboratory specializing in large-scale scientific research involving big data, high-performance computing, and ML. Unlike traditional university settings, National Laboratories achieve scientific breakthroughs by putting together interdisciplinary teams that follow well-devised procedures, with strong division of labor and responsibilities. That kind of research model by production chain was born at Berkeley Lab almost 90 years ago and inspired the meta-strategy paradigm explained in Sections 1.2.2 and 1.3.1.

1.7 ACKNOWLEDGMENTS

Dr. Horst Simon, who is the deputy director of Lawrence Berkeley National Laboratory, accepted to co-author Chapter 22 with Dr. Kesheng Wu, who leads several projects at Berkeley Lab and the National Energy Research Scientific Computing Center (NERSC).³ ML requires extreme amounts of computing power, and my research would not have been possible without their generous support and guidance. In that chapter, Horst and Kesheng explain how Berkeley Lab satisfies the supercomputing needs of researchers worldwide, and the instrumental role played by ML and big data in today's scientific breakthroughs.

Prof. Riccardo Rebonato was the first to read this manuscript and encouraged me to publish it. My many conversations with Prof. Frank Fabozzi on these topics were instrumental in shaping the book in its current form. Very few people in academia have Frank's and Riccardo's industry experience, and very few people in the industry have Riccardo's and Frank's academic pedigree.

Over the past two decades, I have published nearly a hundred works on this book's subject, including journal articles, books, chapters, lectures, source code, etc. In my latest count, these works were co-authored with more than 30 leading experts in this field, including Prof. David H. Bailey (15 articles), Prof. David Easley (8 articles), Prof. Maureen O'Hara (8 articles), and Prof. Jonathan M. Borwein (6 articles). This book is to a great extent also theirs, for it would not have been possible without their support, insights, and continuous exchange of ideas over the years. It would take too long to give them proper credit, so instead I have published the following link where you can find our collective effort: <http://www.quantresearch.org/Co-authors.htm>.

Last but not least, I would like to thank some of my research team members for proofreading the book and helping me produce some of the figures: Diego Aparicio,

³ <http://www.nersc.gov/about>.

Dr. Lee Cohn, Dr. Michael Lewis, Dr. Michael Lock, Dr. Yaxiong Zeng, and Dr. Zhibai Zhang.

EXERCISES

- 1.1 Are you aware of firms that have attempted to transition from discretionary investments to ML-led investments, or blending them into what they call “quantamental” funds?
 - (a) Have they succeeded?
 - (b) What are the cultural difficulties involved in this transition?
- 1.2 What is the most important open problem in mathematical finance? If this problem was resolved, how could:
 - (a) regulators use it to grant investment management licenses?
 - (b) investors use it to allocate funds?
 - (c) firms use it to reward researchers?
- 1.3 According to *Institutional Investor*, only 17% of hedge fund assets are managed by quantitative firms. That is about \$500 billion allocated in total across all quantitative funds as of June 2017, compared to \$386 billion a year earlier. What do you think is driving this massive reallocation of assets?
- 1.4 According to *Institutional Investor*’s Rich List, how many quantitative investment firms are placed within the top 10 most profitable firms? How does that compare to the proportion of assets managed by quantitative funds?
- 1.5 What is the key difference between econometric methods and ML? How would economics and finance benefit from updating their statistical toolkit?
- 1.6 Science has a very minimal understanding of how the human brain (or any brain) works. In this sense, the brain is an absolute black box. What do you think causes critics of financial ML to disregard it as a black box, while embracing discretionary investing?
- 1.7 You read a journal article that describes an investment strategy. In a backtest, it achieves an annualized Sharpe ratio in excess of 2, with a confidence level of 95%. Using their dataset, you are able to reproduce their result in an independent backtest. Why is this discovery likely to be false?
- 1.8 Investment advisors are plagued with conflicts of interest while making decisions on behalf of their investors.
 - (a) ML algorithms can manage investments without conflict of interests. Why?
 - (b) Suppose that an ML algorithm makes a decision that leads to a loss. The algorithm did what it was programmed to do, and the investor agreed to the terms of the program, as verified by forensic examination of the computer logs. In what sense is this situation better for the investor, compared to a loss caused by a discretionary PM’s poor judgment? What is the investor’s recourse in each instance?
 - (c) Would it make sense for financial advisors to benchmark their decisions against the decisions made by such neutral agents?

REFERENCES

- Bailey, D., P. Borwein, and S. Plouffe (1997): “On the rapid computation of various polylogarithmic constants.” *Mathematics of Computation*, Vol. 66, No. 218, pp. 903–913.
- Calkin, N. and M. López de Prado (2014a): “Stochastic flow diagrams.” *Algorithmic Finance*, Vol. 3, No. 1, pp. 21–42.
- Calkin, N. and M. López de Prado (2014b): “The topology of macro financial flows: An application of stochastic flow diagrams.” *Algorithmic Finance*, Vol. 3, No. 1, pp. 43–85.
- Easley, D., M. López de Prado, and M. O’Hara (2013): *High-Frequency Trading*, 1st ed. Risk Books.
- López de Prado, M. (2014): “Quantitative meta-strategies.” *Practical Applications, Institutional Investor Journals*, Vol. 2, No. 3, pp. 1–3.
- López de Prado, M. (2015): “The Future of Empirical Finance.” *Journal of Portfolio Management*, Vol. 41, No. 4, pp. 140–144.
- Stigler, Stephen M. (1981): “Gauss and the invention of least squares.” *Annals of Statistics*, Vol. 9, No. 3, pp. 465–474.

BIBLIOGRAPHY

- Abu-Mostafa, Y., M. Magdon-Ismail, and H. Lin (2012): *Learning from Data*, 1st ed. AMLBook.
- Akansu, A., S. Kulkarni, and D. Malioutov (2016): *Financial Signal Processing and Machine Learning*, 1st ed. John Wiley & Sons-IEEE Press.
- Aronson, D. and T. Masters (2013): *Statistically Sound Machine Learning for Algorithmic Trading of Financial Instruments: Developing Predictive-Model-Based Trading Systems Using TSSB*, 1st ed. CreateSpace Independent Publishing Platform.
- Boyarshinov, V. (2012): *Machine Learning in Computational Finance: Practical Algorithms for Building Artificial Intelligence Applications*, 1st ed. LAP LAMBERT Academic Publishing.
- Cerniglia, J., F. Fabozzi, and P. Kolm (2016): “Best practices in research for quantitative equity strategies.” *Journal of Portfolio Management*, Vol. 42, No. 5, pp. 135–143.
- Chan, E. (2017): *Machine Trading: Deploying Computer Algorithms to Conquer the Markets*, 1st ed. John Wiley & Sons.
- Gareth, J., D. Witten, T. Hastie, and R. Tibshirani (2013): *An Introduction to Statistical Learning: with Applications in R*, 1st ed. Springer.
- Geron, A. (2017): *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st ed. O’Reilly Media.
- Gyorfi, L., G. Ottucsak, and H. Walk (2012): *Machine Learning for Financial Engineering*, 1st ed. Imperial College Press.
- Hackeling, G. (2014): *Mastering Machine Learning with Scikit-Learn*, 1st ed. Packt Publishing.
- Hastie, T., R. Tibshirani, and J. Friedman (2016): *The Elements of Statistical Learning*, 2nd ed. Springer-Verlag.
- Hauck, T. (2014): *Scikit-Learn Cookbook*, 1st ed. Packt Publishing.
- McNelis, P. (2005): *Neural Networks in Finance*, 1st ed. Academic Press.
- Raschka, S. (2015): *Python Machine Learning*, 1st ed. Packt Publishing.

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, 52*f*
 - measurement of, 206
- AdaBoost implementation, 100, 100*f*
- Adaptable I/O System (ADIOS), 336–337, 339, 340
- Alternative data, 24*t*, 25
- Amihud’s lambda, 288–289, 289*f*
- Analytics, 24*t*, 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, 225*f*, 229, 231*f*, 232, 234
 - diversification in, 4, 9, 222, 224, 234, 238
 - Markowitz’s approach to, 221–222
 - Monte Carlo simulations for, 234–236, 235*f*–236*f*, 242–244
 - numerical example of, 231–233, 232*f*, 233*f*, 233*t*
 - practical problems in, 222–223, 223*f*
 - quasi-diagonalization in, 224, 229, 233*f*
 - recursive bisection in, 224, 229–231
 - risk-based, 222. *See also* Risk-based asset allocation approaches
 - tree clustering approaches to, 224–229, 225*f*, 228*f*, 232*f*
- 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, 157*f*
 - 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
 - historical simulations in trading rules and, 170–172, 178–179, 187
 - 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, 202*f*
 - 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, 97*f*
 - 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, 95*f*
- Bars (table rows), 25–32
- categories of, 26
 - dollar bars, 27–28, 28*f*, 44
 - dollar imbalance bars, 29–30
 - dollar runs bars, 31–32

- 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, 143^f
 - runs and increase in, 199
 - size discretization in, 144–145, 145^f
 - strategy-independent approaches to, 141–142
 - strategy's capacity and, 196
- Bet timing, deriving, 197
- Betting frequency
 - backtesting and, 196
 - computing, 215–216, 216^f
 - implied precision computation and, 214–215, 215^f
 - investment strategy with trade-off between precision and, 212–213, 212^f
 - 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, 237^f, 330, 331–332, 340
- Bloomberg, 23, 36
- Boosting, 99–100
 - AdaBoost implementation of, 100, 100^f
 - 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, 164^f

- 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, 256*f*, 257*f*
- 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, 235*f*–236*f*, 242–244
 - numerical example with, 231–233, 232*f*, 233*f*, 233*t*
 - open-source implementation of, 222
 - practical problems with, 222–223, 223*f*
- Cross-entropy loss (log loss) scoring, 133–134, 135*f*
- 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, 108*f*
 - failures in finance using, 104
 - goal of, 103
 - hyper-parameter tuning with, 129–135
 - k-fold, 103–109, 104*f*
 - 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, 107*f*
 - sklearn bugs in, 109–110
- CUSUM filter, 38–40, 40*f*
- 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, 205*f*
- Deployment team, 8
- Dickey-Fuller test
 - Chow type, 251–252
 - supremum augmented (SADF), 252–259, 253*f*, 257*f*
- Discretionary portfolio managers (PMs), 4–5, 15
- Discretization of bet size, 144–145, 145*f*
- Diversification, 4, 9, 222, 224, 234, 238
- Dollar bars, 27–28, 28*f*, 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, 202*f*
 - run measurements using, 202
- Dynamic bet sizes, 145–148

- Econometrics, 14, 85
 financial Big Data analysis and, 236
 financial machine learning versus, 15
 HRP approach compared with, 236, 237*f*
 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*f*
 information ratio and, 205
 probabilistic Sharpe ratio (PSR) and, 203–204, 204*f*, 205–206, 218
 Sharpe ratio (SR) definition in, 203
- Efficient frontier, 222
- Electricity consumption analysis, 340–341, 342*f*–343*f*
- 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, 274*f*
 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, 40*f*
- Excess returns, in information ratio, 205
- Execution costs, 202–203
- Expanding window method, 80–82, 81*f*
- Explosiveness tests, 249, 251–259
- Chow-type Dickey-Fuller test, 251–252
 supremum augmented Dickey-Fuller (SADF) test, 252–259, 253*f*, 257*f*
- 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, 24*t*
 fundamental, 23–24
 market, 24–25
- Financial data structures, 23–40
 bars (table rows) in, 25–32

- 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, 83*f*
- 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
- Flash Crash of 2010, 296, 329–330, 341–345
- 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, 81*f*
 fixed-width window fracdiff (FFD) method for, 82–84, 83*f*
 maximum memory preservation in, 84–85, 84*f*, 86*t*–87*t*
- Frequency. *See* Betting frequency
- Fundamental data, 23–24, 24*t*
- Fusion collaboration project, 338–340, 339*f*
- 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, 290*f*
- 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, 237*f*
 full implementation of, 240–242
 Monte Carlo simulations using, 234–236, 235*f*–236*f*, 242–244
 numerical example of, 231–233, 232*f*, 233*f*, 233*t*
 practical application of, 221
 quasi-diagonalization in, 224, 229
 recursive bisection in, 224, 229–231
 standard approaches compared with, 221, 236–238

- traditional risk parity approach
 compared with, 231–232, 233*t*,
 234, 235*f*
 tree clustering approaches to,
 224–229, 225*f*, 228*f*, 232*f*
 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*f*–343*f*
 Flash Crash of 2010 response and,
 329–330, 341–345
 fusion collaboration project using,
 338–340, 339*f*
 global dynamic optimal trajectory
 and, 325–327
 hardware for, 331–335, 332*f*, 333*f*,
 334*f*
 integer optimization approach and,
 321–325, 322*f*
 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*f*
 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*f*
 randomized search cross-validation
 and, 131–133

 Implementation shortfall statistics,
 202–203
 Implied betting frequency, 215–216,
 216*f*
 Implied precision computation,
 214–215, 215*f*
 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, 232*f*, 233*f*, 233*t*
 Monte Carlo simulations using,
 234–236, 235*f*–236*f*, 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*f*

- 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, 212*f*
 - 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, 104*f*
 - embargo on training observations in, 107–108, 108*f*
 - 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, 107*f*
 - when used, 104
- Kyle's lambda, 286–288, 288*f*
- 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, 47*f*
- Labels
 - average uniqueness over lifespan of, 61–62, 61*f*
 - 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, 107*f*
 - 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, 135*f*
- 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, 235*f*–236*f*, 242–244
 - numerical example of, 231–233, 232*f*, 233*f*, 233*t*
 - quasi-diagonalization in, 224, 229, 233*f*

- recursive bisection in, 224, 229–231
 - tree clustering approaches to, 224–229, 225*f*, 228*f*, 232*f*
- Market data, 24–25, 24*t*
- 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, 126*f*
 - 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, 125*f*
 - 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, 289*f*
 - bid-ask spread estimator (Corwin-Schultz algorithm) and, 284–286
 - Hasbrouck’s lambda and, 289, 290*f*
 - high-low volatility estimator and, 283–284
 - Kyle’s lambda and, 286–288, 288*f*
 - 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, 235*f*–236*f*, 242–244
 - sequential bootstraps evaluation using, 66–68, 68*f*
- Multi-product series, 32–37
 - ETF trick for, 33–34
 - PCA weights for, 35–36, 35*f*
 - 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, 135*f*
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, 186*f*,
187*f*–191*f*
cases with positive long-run
equilibrium in, 180–182, 181*f*,
182*f*, 183*f*–186*f*
cases with zero long-run equilibrium
in, 177–180, 177*f*, 178*f*, 179*f*
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, 225*f*,
229, 231*f*, 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, 235*f*–236*f*, 242–244
numerical example of, 231–233, 232*f*,
233*f*, 233*t*
practical problems in, 222–223, 223*f*
tree clustering approaches to,
224–229, 225*f*, 228*f*, 232*f*
- 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, 52*f*
investment strategy with trade-off
between frequency and, 212–213,
212*f*
measurement of, 206
- Predicted probabilities, in bet sizing,
142–144, 143*f*

- Principal components analysis (PCA)
 hedging weights using, 35–36, 35*f*
 linear substitution effects and, 118, 119–121
- Probabilistic Sharpe ratio (PSR)
 calculation of, 203–204, 204*f*
 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, 157*f*, 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, 216–217
- 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, 186*f*, 187*f*–191*f*
 cases with positive long-run equilibrium and, 180–182, 181*f*, 182*f*, 183*f*–186*f*
 cases with zero long-run equilibrium and, 177–180, 177*f*, 178*f*, 179*f*
 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
 triple-barrier labeling method for, 45–46, 47*f*
- Purged K-fold cross-validation (CV)
 grid search cross-validation and, 129–131
 hyper-parameter tuning with, 129–135
 implementation of, 105–106, 107*f*
 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, 52*f*
 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, 35*f*
 - 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, 233*t*, 234, 235*f*
- Rolled prices, 37
- Roll model, 282–283

- Sample weights, 59–72
 - average uniqueness of labels over lifespan and, 61–62, 61*f*
 - 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, 72*f*
- Sampling features, 38–40
 - downsampling and, 38
 - event-based sampling and, 38–40, 40*f*

- 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, 68*f*
 - numerical example of, 65–66
- Shannon, Claude, 263–264
- Sharpe ratio (SR) in efficiency measurements
 - annualized, 205
 - definition of, 203
 - deflated (DSR), 204, 205*f*
 - information ratio and, 205
 - probabilistic (PSR), 203–204, 204*f*, 205–206, 218
 - purpose of, 203
 - targeting, for various betting frequencies, 212–213
- Shorting, in quantitative investing, 152

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

- 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, 72*f*
- 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
 example of, 202*f*
 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, 225*f*, 228*f*, 232*f*
- Triple-barrier labeling method, 45–46, 47*f*, 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

(continued from front flap)

Stop guessing and profit off data by:

- Tackling today's most challenging aspects of applying ML algorithms to financial strategies, including backtest overfitting
- Using improved tactics to structure financial data so it produces better outcomes with ML algorithms
- Conducting superior research with ML algorithms, as well as accurately validating the solutions you discover
- Learning the tricks of the trade from one of the largest ML investment managers

Put yourself ahead of tomorrow's competition today with *Advances in Financial Machine Learning*.



Photo by Sarah Griffin/WireImage

DR. MARCOS LÓPEZ DE PRADO manages several multibillion-dollar funds for institutional investors using ML algorithms. Marcos is also a research fellow at Lawrence Berkeley National Laboratory (U.S. Department of Energy, Office of Science). One of the top-10 most read authors in finance (SRRN's rankings), he has published dozens of scientific articles on ML in the leading academic journals, and he holds multiple international patent applications on algorithmic trading. Marcos earned a PhD in Financial Economics (2003), a second PhD in Mathematical Finance (2011) from Universidad Complutense de Madrid, and is a recipient of Spain's National Award for Academic Excellence (1999). He completed his postdoctoral research at Harvard University and Cornell University, where he teaches a Financial ML course at the School of Engineering. Marcos has an Erdős #2 and an Einstein #4 according to the American Mathematical Society.

LOPEZ DE PRADO

MARCOS LÓPEZ DE PRADO

Praise for

ADVANCES in FINANCIAL MACHINE LEARNING

"Dr. López de Prado has written the first comprehensive book describing the application of modern ML to financial modeling. The book blends the latest technological developments in ML with critical life lessons learned from the author's decades of financial experience in leading academic and industrial institutions. I highly recommend this exciting book to both prospective students of financial ML and the professors and supervisors who teach and guide them."

—**PROF. PETER CARR**, Chair of the Finance and Risk Engineering Department, NYU Tandon School of Engineering

"Financial problems require very distinct machine learning solutions. Dr. López de Prado's book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book."

—**PROF. FRANK FABOZZI**, EDHEC Business School, Editor of *The Journal of Portfolio Management*

"Marcos has assembled in one place an invaluable set of lessons and techniques for practitioners seeking to deploy machine learning methods in finance. Marcos's insightful book is laden with useful advice to help keep a curious practitioner from going down any number of blind alleys, or shooting oneself in the foot."

—**ROSS GARON**, Head of Cubist Systematic Strategies, Managing Director, Point72 Asset Management

"The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine learning is the second wave and it will touch every aspect of finance. López de Prado's *Advances in Financial Machine Learning* is essential for readers who want to be ahead of the technology rather than being replaced by it."

—**PROF. CAMPBELL HARVEY**, Duke University, Former President of the American Finance Association

"The author's academic and professional finance credentials shine through the pages of this book — indeed, I could think of few, if any, authors better suited to explaining both the theoretical and the practical aspects of this new and (for most) unfamiliar subject. Destined to become a classic in this rapidly burgeoning field."

—**PROF. RICCARDO REBONATO**, EDHEC Business School, Former Global Head of Rates and FX Analytics at PIMCO

Cover Design: Wiley

ISBN: 978-1-119-48208-6



9 781119 482086

Subscribe to our free Finance and Investing eNewsletter at wiley.com/enewletters. Visit wileyfinance.com

WILEY



WILEY

ADVANCES in FINANCIAL MACHINE LEARNING



Today's machine learning (ML) algorithms have conquered the major strategy games, and are routinely used to execute tasks once only possible by a limited group of experts. Over the next few years, ML algorithms will transform finance beyond anything we know today. *Advances in Financial Machine Learning* was written for the investment professionals and data scientists at the forefront of this evolution.

This one-of-a-kind, practical guidebook is your go-to resource of authoritative insight into using advanced ML solutions to overcome real-world investment problems. It demystifies the entire subject and unveils cutting-edge ML techniques specific to investing. With step-by-step clarity and purpose, it quickly brings you up to speed on fully proven approaches to data analysis, model research, and discovery evaluation. Then, it shines a light on the nuanced details behind innovative ways to extract informative features from financial data. To streamline implementation, it gives you valuable recipes for high-performance computing systems optimized to handle this type of financial data analysis.

Advances in Financial Machine Learning crosses the proverbial divide that separates academia and the industry. It does not advocate a theory merely because of its mathematical beauty, and it does not propose a solution just because it appears to work. The author transmits the kind of knowledge that only comes from experience, formalized in a rigorous manner.

This turnkey guide is designed to be immediately useful to the practitioner, by featuring code snippets and hands-on exercises that facilitate the quick absorption and application of best practices in the real world.

(continued on back flap)