Table of Contents

Preface xiii

Chapter 1: Machine Learning for Trading – From Idea to Execution 1

The rise of ML in the investment industry 2

From electronic to high-frequency trading 3

Factor investing and smart beta funds 5

Algorithmic pioneers outperform humans 7

ML and alternative data 10

Crowdsourcing trading algorithms 11

Designing and executing an ML-driven strategy 12

Sourcing and managing data 13

From alpha factor research to portfolio management 13

Strategy backtesting 15

ML for trading – strategies and use cases 15

The evolution of algorithmic strategies 15

Use cases of ML for trading 16

Summary 19

Chapter 2: Market and Fundamental Data – Sources and Techniques 21

Market data reflects its environment 22

Market microstructure – the nuts and bolts 23

How to trade – different types of orders 23

Where to trade – from exchanges to dark pools 24

Working with high-frequency data 26

How to work with Nasdaq order book data 26

Communicating trades with the FIX protocol 27

The Nasdaq TotalView-ITCH data feed 27

From ticks to bars – how to regularize market data 35

AlgoSeek minute bars – equity quote and trade data 40

API access to market data 44

Remote data access using pandas 44

yfinance – scraping data from Yahoo! Finance 46

Quantopian 48
Zipline 48
Quandl 50
Other market data providers 50
How to work with fundamental data 51
Financial statement data 51
Other fundamental data sources 56
Efficient data storage with pandas 57
Summary 58
Chapter 3: Alternative Data for Finance – Categories and Use Cases 59
The alternative data revolution 60
Sources of alternative data 62
Individuals 62
Business processes 63
Sensors 63
Criteria for evaluating alternative data 65
Quality of the signal content 65
Quality of the data 67
Technical aspects 68
The market for alternative data 69
Data providers and use cases 70
Working with alternative data 72
Scraping OpenTable data 72
Scraping and parsing earnings call transcripts 77
Summary 80
Chapter 4: Financial Feature Engineering – How to Research Alpha Factors 81
Alpha factors in practice – from data to signals 82
Building on decades of factor research 84
Momentum and sentiment – the trend is your friend 84
Value factors – hunting fundamental bargains 88
Volatility and size anomalies 90
Quality factors for quantitative investing 92
Engineering alpha factors that predict returns 94
How to engineer factors using pandas and NumPy 94
How to use TA-Lib to create technical alpha factors 99
Denoising alpha factors with the Kalman filter 100
How to preprocess your noisy signals using wavelets 104
From signals to trades – Zipline for backtests 106
How to backtest a single-factor strategy 106
Combining factors from diverse data sources 109
Separating signal from noise with Alphalens 111
Creating forward returns and factor quantiles 112
Predictive performance by factor quantiles 113

The information coefficient 115
Factor turnover 117
Alpha factor resources 118
Alternative algorithmic trading libraries 118
Summary 119
Chapter 5: Portfolio Optimization and Performance Evaluation 121
How to measure portfolio performance 122
Capturing risk-return trade-offs in a single number 122
The fundamental law of active management 124
How to manage portfolio risk and return 125
The evolution of modern portfolio management 125
Mean-variance optimization 127
Alternatives to mean-variance optimization 131
Risk parity 134
Risk factor investment 135
Hierarchical risk parity 135
Trading and managing portfolios with Zipline 136
Scheduling signal generation and trade execution 137
Implementing mean-variance portfolio optimization 138
Measuring backtest performance with pyfolio 140
Creating the returns and benchmark inputs 141
Walk-forward testing – out-of-sample returns 142
Summary 146
Chapter 6: The Machine Learning Process 147
How machine learning from data works 148
The challenge – matching the algorithm to the task 149
Supervised learning – teaching by example 149
Unsupervised learning – uncovering useful patterns 150
Reinforcement learning – learning by trial and error 152
The machine learning workflow 153
Basic walkthrough – k-nearest neighbors 154
Framing the problem – from goals to metrics 154
Collecting and preparing the data 160
Exploring, extracting, and engineering features 160
Selecting an ML algorithm 162
Design and tune the model 162
How to select a model using cross-validation 165
How to implement cross-validation in Python 166
Challenges with cross-validation in finance 168
Parameter tuning with scikit-learn and Yellowbrick 170
Summary 172
Chapter 7: Linear Models – From Risk Factors to Return Forecasts 173
From inference to prediction 174

The baseline model – multiple linear regression **175**
How to formulate the model 175
How to train the model 176
The Gauss–Markov theorem 179
How to conduct statistical inference 180
How to diagnose and remedy problems 181
How to run linear regression in practice **184**
OLS with statsmodels 184
Stochastic gradient descent with sklearn 186
How to build a linear factor model **187**
From the CAPM to the Fama–French factor models 188
Obtaining the risk factors 189
Fama–Macbeth regression 191
Regularizing linear regression using shrinkage **194**
How to hedge against overfitting 194
How ridge regression works 195
How lasso regression works 196
How to predict returns with linear regression **197**
Preparing model features and forward returns 197
Linear OLS regression using statsmodels 203
Linear regression using scikit-learn 205
Ridge regression using scikit-learn 208
Lasso regression using sklearn 210
Comparing the quality of the predictive signals 212
Linear classification **212**
The logistic regression model 213
How to conduct inference with statsmodels 215
Predicting price movements with logistic regression 217
Summary **219**
Chapter 8: The ML4T Workflow – From Model to Strategy Backtesting **221**
How to backtest an ML-driven strategy **222**
Backtesting pitfalls and how to avoid them **223**
Getting the data right 224
Getting the simulation right 225
Getting the statistics right 226
How a backtesting engine works **227**
Vectorized versus event-driven backtesting 228
Key implementation aspects 230
backtrader – a flexible tool for local backtests **232**
Key concepts of backtrader's Cerebro architecture 232
How to use backtrader in practice 235
backtrader summary and next steps 239
Zipline – scalable backtesting by Quantopian **239**

Calendars and the Pipeline for robust simulations 240
Ingesting your own bundles with minute data 242
The Pipeline API – backtesting an ML signal 245
How to train a model during the backtest 250
Instead of How to use 254
Summary 254
Chapter 9: Time-Series Models for Volatility Forecasts and Statistical Arbitrage 255
Tools for diagnostics and feature extraction 256
How to decompose time-series patterns 257
Rolling window statistics and moving averages 258
How to measure autocorrelation 259
How to diagnose and achieve stationarity 260
Transforming a time series to achieve stationarity 261
Handling instead of How to handle 261
Time-series transformations in practice 263
Univariate time-series models 265
How to build autoregressive models 266
How to build moving-average models 267
How to build ARIMA models and extensions 268
How to forecast macro fundamentals 270
How to use time-series models to forecast volatility 272
Multivariate time-series models 276
Systems of equations 277
The vector autoregressive (VAR) model 277
Using the VAR model for macro forecasts 278
Cointegration – time series with a shared trend 281
The Engle-Granger two-step method 282
The Johansen likelihood-ratio test 282
Statistical arbitrage with cointegration 283
How to select and trade comoving asset pairs 283
Pairs trading in practice 285
Preparing the strategy backtest 288
Backtesting the strategy using backtrader 292
Extensions – how to do better 294
Summary 294
Chapter 10: Bayesian ML – Dynamic Sharpe Ratios and Pairs Trading 295
How Bayesian machine learning works 296
How to update assumptions from empirical evidence 297
Exact inference – maximum a posteriori estimation 298
Deterministic and stochastic approximate inference 301
Probabilistic programming with PyMC3 305
Bayesian machine learning with Theano 305

The PyMC3 workflow: predicting a recession 305

Bayesian ML for trading 317

Bayesian Sharpe ratio for performance comparison 317

Bayesian rolling regression for pairs trading 320

Stochastic volatility models 323

Summary 326

Chapter 11: Random Forests – A Long-Short Strategy for Japanese Stocks 327

Decision trees – learning rules from data 328

How trees learn and apply decision rules 328

Decision trees in practice 330

Overfitting and regularization 336

Hyperparameter tuning 338

Random forests – making trees more reliable 345

Why ensemble models perform better 345

Bootstrap aggregation 346

How to build a random forest 349

How to train and tune a random forest 350

Feature importance for random forests 352

Out-of-bag testing 352

Pros and cons of random forests 353

Long-short signals for Japanese stocks 353

The data – Japanese equities 354

The ML4T workflow with LightGBM 355

The strategy – backtest with Zipline 362

Summary 364

Chapter 12: Boosting Your Trading Strategy 365

Getting started – adaptive boosting 366

The AdaBoost algorithm 367

Using AdaBoost to predict monthly price moves 368

Gradient boosting – ensembles for most tasks 370

How to train and tune GBM models 372

How to use gradient boosting with sklearn 374

Using XGBoost, LightGBM, and CatBoost 378

How algorithmic innovations boost performance 379

A long-short trading strategy with boosting 383

Generating signals with LightGBM and CatBoost 383

Inside the black box - interpreting GBM results 391

Backtesting a strategy based on a boosting ensemble 399

Lessons learned and next steps 401

Boosting for an intraday strategy 402

Engineering features for high-frequency data 402

Minute-frequency signals with LightGBM 404

Evaluating the trading signal quality 405

Summary 406
Chapter 13: Data-Driven Risk Factors and Asset Allocation with Unsupervised Learning 407
Dimensionality reduction 408
The curse of dimensionality 409
Linear dimensionality reduction 411
Manifold learning – nonlinear dimensionality reduction 418
PCA for trading 421
Data-driven risk factors 421
Eigenportfolios 424
Clustering 426
k-means clustering 427
Hierarchical clustering 429
Density-based clustering 431
Gaussian mixture models 432
Hierarchical clustering for optimal portfolios 433
How hierarchical risk parity works 433
Backtesting HRP using an ML trading strategy 435
Summary 438
Chapter 14: Text Data for Trading – Sentiment Analysis 439
ML with text data – from language to features 440
Key challenges of working with text data 440
The NLP workflow 441
Applications 443
From text to tokens – the NLP pipeline 443
NLP pipeline with spaCy and textacy 444
NLP with TextBlob 448
Counting tokens – the document-term matrix 449
The bag-of-words model 450
Document-term matrix with scikit-learn 451
Key lessons instead of lessons learned 455
NLP for trading 455
The naive Bayes classifier 456
Classifying news articles 457
Sentiment analysis with Twitter and Yelp data 458
Summary 462
Chapter 15: Topic Modeling – Summarizing Financial News 463
Learning latent topics – Goals and approaches 464
Latent semantic indexing 465
How to implement LSI using sklearn 466
Strengths and limitations 468
Probabilistic latent semantic analysis 469
How to implement pLSA using sklearn 470

Strengths and limitations 471

Latent Dirichlet allocation 471

How LDA works 471

How to evaluate LDA topics 473

How to implement LDA using sklearn 475

How to visualize LDA results using pyLDAvis 475

How to implement LDA using Gensim 476

Modeling topics discussed in earnings calls 478

Data preprocessing 478

Model training and evaluation 479

Running experiments 480

Topic modeling for with financial news 481

Summary 482

Chapter 16: Word Embeddings for Earnings Calls and SEC Filings 483

How word embeddings encode semantics 484

How neural language models learn usage in context 485

word2vec – scalable word and phrase embeddings 485

Evaluating embeddings using semantic arithmetic 487

How to use pretrained word vectors 489

GloVe – Global vectors for word representation 489

Custom embeddings for financial news 491

Preprocessing – sentence detection and n-grams 492

The skip-gram architecture in TensorFlow 2 493

Visualizing embeddings using TensorBoard 496

How to train embeddings faster with Gensim 497

word2vec for trading with SEC filings 499

Preprocessing – sentence detection and n-grams 500

Model training 501

Sentiment analysis using doc2vec embeddings 503

Creating doc2vec input from Yelp sentiment data 503

Training a doc2vec model 504

Training a classifier with document vectors 505

Lessons learned and next steps 507

New frontiers – pretrained transformer models 507

Attention is all you need 508

BERT – towards a more universal language model 509

Trading on text data – lessons learned and next steps 511

Summary 511

Chapter 17: Deep Learning for Trading 513

Deep learning – what's new and why it matters 514

Hierarchical features tame high-dimensional data 515

DL as representation learning 516

How DL relates to ML and AI 517

Designing an NN 518

A simple feedforward neural network architecture 519
Key design choices 520
How to regularize deep NNs 522
Training faster – optimizations for deep learning 523
Summary – how to tune key hyperparameters 525
A neural network from scratch in Python 526
The input layer 526
The hidden layer 527
The output layer 528
Forward propagation 529
The cross-entropy cost function 529
How to implement backprop using Python 529
Popular deep learning libraries 534
Leveraging GPU acceleration 534
How to use TensorFlow 2 535
How to use TensorBoard 537
How to use PyTorch 1.4 538
Alternative options 541
Optimizing an NN for a long-short strategy 542
Engineering features to predict daily stock returns 542
Defining an NN architecture framework 542
Cross-validating design options to tune the NN 543
Evaluating the predictive performance 545
Backtesting a strategy based on ensembled signals 547
How to further improve the results 549
Summary 549
Chapter 18: CNNs for Financial Time Series and Satellite Images 551
How CNNs learn to model grid-like data 552
From hand-coding to learning filters from data 553
How the elements of a convolutional layer operate 554
The evolution of CNN architectures: key innovations 558
CNNs for satellite images and object detection 559
LeNet5 – The first CNN with industrial applications 560
AlexNet – reigniting deep learning research 563
Transfer learning – faster training with less data 565
Object detection and segmentation 573
Object detection in practice 573
CNNs for time-series data – predicting returns 577
An autoregressive CNN with 1D convolutions 577
CNN-TA – clustering time series in 2D format 581
Summary 589
Chapter 19: RNNs for Multivariate Time Series and Sentiment Analysis 591
How recurrent neural nets work 592

Unfolding a computational graph with cycles 594
Backpropagation through time 594
Alternative RNN architectures 595
How to design deep RNNs 596
The challenge of learning long-range dependencies 597
Gated recurrent units 599
RNNs for time series with TensorFlow 2 599
Univariate regression – predicting the S&P 500 600
How to get time series data into shape for an RNN 600
Stacked LSTM – predicting price moves and returns 605
Multivariate time-series regression for macro data 611
RNNs for text data 614
LSTM with embeddings for sentiment classification 614
Sentiment analysis with pretrained word vectors 617
Predicting returns from SEC filing embeddings 619
Summary 624
Chapter 20: Autoencoders for Conditional Risk Factors and Asset Pricing 625
Autoencoders for nonlinear feature extraction 626
Generalizing linear dimensionality reduction 626
Convolutional autoencoders for image compression 627
Managing overfitting with regularized autoencoders 628
Fixing corrupted data with denoising autoencoders 628
Seq2seq autoencoders for time series features 629
Generative modeling with variational autoencoders 629
Implementing autoencoders with TensorFlow 2 630
How to prepare the data 630
One-layer feedforward autoencoder 631
Feedforward autoencoder with sparsity constraints 634
Deep feedforward autoencoder 634
Convolutional autoencoders 636
Denoising autoencoders 637
A conditional autoencoder for trading 638
Sourcing stock prices and metadata information 639
Computing predictive asset characteristics 641
Creating the conditional autoencoder architecture 643
Lessons learned and next steps 648
Summary 648
Chapter 21: Generative Adversarial Networks for Synthetic Time-Series Data 649
Creating synthetic data with GANs 650
Comparing generative and discriminative models 651
Adversarial training – a zero-sum game of trickery 651
The rapid evolution of the GAN architecture zoo 652

GAN applications to images and time-series data 653
How to build a GAN using TensorFlow 2 655
Building the generator network 655
Creating the discriminator network 656
Setting up the adversarial training process 657
Evaluating the results 660
TimeGAN for synthetic financial data 660
Learning to generate data across features and time 661
Implementing TimeGAN using TensorFlow 2 663
Evaluating the quality of synthetic time-series data 672
Lessons learned and next steps 678
Summary 678
Chapter 22: Deep Reinforcement Learning – Building a Trading Agent 679
Elements of a reinforcement learning system 680
The policy – translating states into actions 681
Rewards – learning from actions 681
The value function – optimal choice for the long run 682
With or without a model – look before you leap? 682
How to solve reinforcement learning problems 682
Key challenges in solving RL problems 683
Fundamental approaches to solving RL problems 683
Solving dynamic programming problems 684
Finite Markov decision problems 684
Policy iteration 687
Value iteration 688
Generalized policy iteration 688
Dynamic programming in Python 689
Q-learning – finding an optimal policy on the go 694
Exploration versus exploitation – ε-greedy policy 695
The Q-learning algorithm 695
How to train a Q-learning agent using Python 695
Deep RL for trading with the OpenAI Gym 696
Value function approximation with neural networks 697
The Deep Q-learning algorithm and extensions 697
Introducing the OpenAI Gym 699
How to implement DDQN using TensorFlow 2 700
Creating a simple trading agent 704
How to design a custom OpenAI trading environment 705
Deep Q-learning on the stock market 709
Lessons learned 711
Summary 711
Chapter 23: Conclusions and Next Steps 713
Key takeaways and lessons learned 714

Data is the single most important ingredient 715
Domain expertise – telling the signal from the noise 716
ML is a toolkit for solving problems with data 717
Beware of backtest overfitting 719
How to gain insights from black-box models 719
ML for trading in practice 720
Data management technologies 720
ML tools 722
Online trading platforms 722
Conclusion 723
Appendix: Alpha Factor Library 725
Common alpha factors implemented in TA-Lib 726
A key building block – moving averages 726
Overlap studies – price and volatility trends 729
Momentum indicators 733
Volume and liquidity indicators 741
Volatility indicators 743
Fundamental risk factors 744
WorldQuant's quest for formulaic alphas 745
Cross-sectional and time-series functions 745
Formulaic alpha expressions 747
Bivariate and multivariate factor evaluation 749
Information coefficient and mutual information 749
Feature importance and SHAP values 750
Comparison – the top 25 features for each metric 750
Financial performance – Alphalens 752
References 753
Index 769