Machine learning for trading

Chenxi Yang, Li Yi, Hao Shi

Wuhan University

October 6, 2022

Chapter 1 Machine learning for trading – From Idea to Execution

Chapter 2 Market and Fundamental Data

The author——Stefan Jansen

- Stefan Jansen is the founder and CEO of Applied AI.
- He advises Fortune 500 companies, investment firms, and start-ups across industries on data and AI strategy, building data science teams, and developing end-to-end machine learning solutions.
- He was a partner and managing director at an international investment firm.

What to learn in this book:

- The examples in this book will illustrate how ML algorithms can extract information from data to support or automate key investment activities.
- It involves:
  - Observing the market
  - Analyzing data to form expectations about the future *decide on placing buy or sell orders
  - Managing the resulting portfolio to produce attractive returns
- Final goal: to generate alpha with good information ratio

Factor investing

- The return provided by an asset is a function of the uncertainty or risk associated with the investment.
- Modern portfolio theory (MPT) introduced the distinction between idiosyncratic and systematic sources of risk for a given asset.
- The capital asset pricing model (CAPM) identified a single factor driving all asset returns: the return on the market portfolio in excess of T-bills.

Factor investment

- Numerous additional risk factors have since been discovered
- These risk factors were labeled anomalies since they contradicted the efficient market hypothesis (EMH).
- Such as:size effect, value effect, momentum effectilliquidity premium……

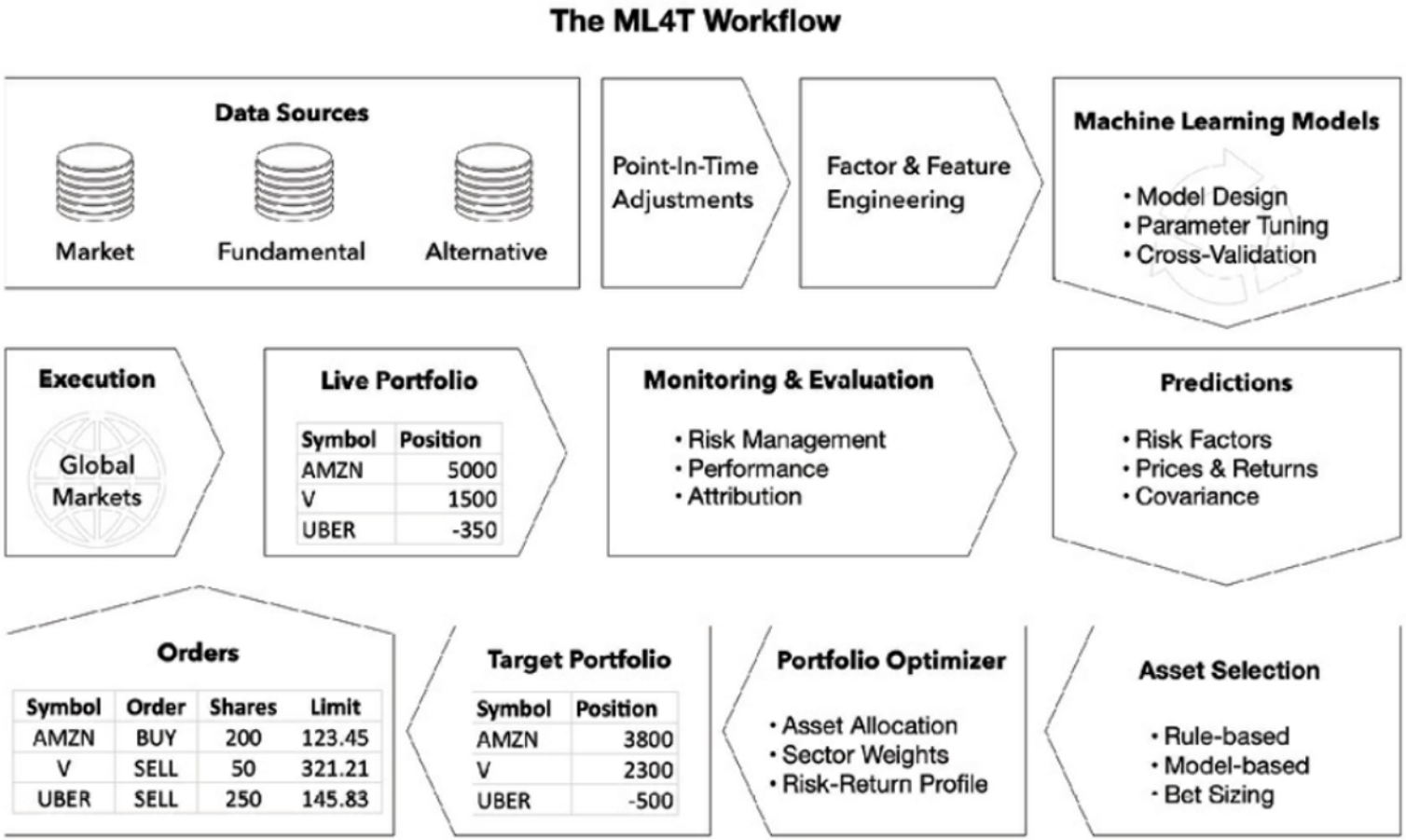
Multifactor models

- Multifactor models define risks in broader and more diverse terms than just the market portfolio.
- Arbitrage pricing theory(APT)
- FF3 (R/SMB/HML)
- FF5(R/SMB/HML/RMW/CMA)

Why ML

- Making precise predictions to generate alpha requires superior information, either through access to better data, a superior ability to process it, or both.
- This is where ML comes in: ML for trading (ML4T) typically aim to make more efficient use of a rapidly diversifying range of data to produce both better and more actionable forecasts, thus improving the quality of investment decisions and results.

Designing and executing an ML-driven strategy

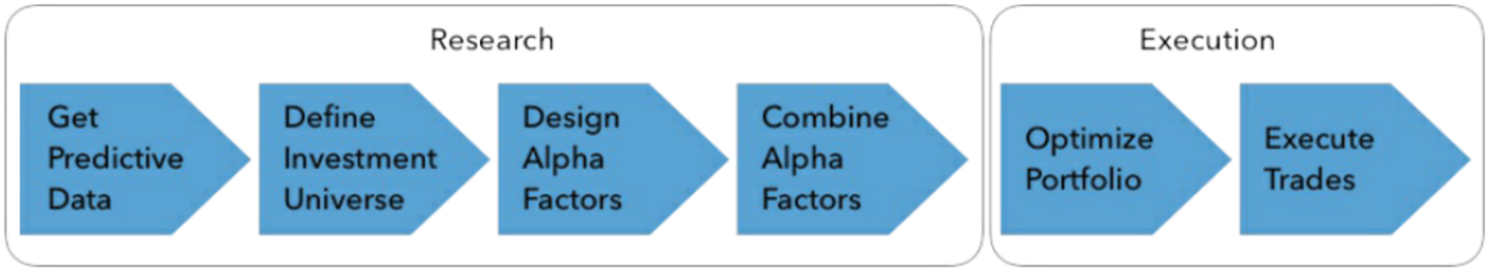


Sourcing and managing data

- The dramatic evolution of data availability in terms of volume, variety, and velocity -careful selection and management:
  - Identify and evaluate market, fundamental, and alternative data sources containing alpha signals.
  - Deploy or access a cloud-based scalable data infrastructure and analytical tools
  - Carefully manage and curate data to avoid look-ahead bias by adjusting it to the desired frequency on a point-in-time basis

From alpha factor research to portfolio management

Alpha factors are designed to extract signals from data to predict returns for a given investment universe over the trading horizon.



The research phase

- The research phase includes the design and evaluation of alpha factors.
- A predictive factor captures some aspect of a systematic relationship between a data source and an important strategy input like asset returns.
- Optimizing the predictive power requires creative feature engineering in the form of effective data transformations.

The execution phase

- During the execution phase, alpha factors emit signals that lead to buy or sell orders.
- The resulting portfolio holdings, in turn, have specific risk profiles that interact and contribute to the aggregate portfolio risk. Portfolio management involves optimizing position sizes to achieve a balance of return and risk of the portfolio that aligns with the investment objectives.

Use cases of ML for trading

- It can be applied at several steps of the trading process:
  - Data mining to identify patterns, extract features, and generate insights
  - Supervised learning to generate risk factors or alphas and create trade ideas
  - The aggregation of individual signals into a strategy
  - The allocation of assets according to risk profiles learned by an algorithm
  - The testing and evaluation of strategies, including through the use of synthetic data
  - The interactive, automated refinement of a strategy using reinforcement learning

Use cases of ML for trading

- Data mining for feature extraction and insights
  - The cost-effective evaluation of large, complex datasets requires the detection of signals at scale
- Supervised learning for alpha factor creation
  - The most familiar rationale for applying ML to trading is to obtain predictions of asset fundamentals, price movements, or market conditions
  - ML predictions can also target specific risk factors, such as value or volatility

Use cases of ML for trading

- Asset allocation
  - ML has been used to allocate portfolios based on decision-tree models that compute a hierarchical form of risk parity.
- Testing trade ideas
  - We will demonstrate various methods to test ML models using market, fundamental, and alternative data sources that obtain sound estimates of out-of-sample errors.
- aims to train agents to learn a policy function based on rewards

The emergence of quantamental funds

- Two distinct approaches have evolved in active investment management: systematic (or quant) and discretionary investing.
  - Systematic approaches rely on algorithms for a repeatable and datadriven approach
  - A discretionary approach involves an in-depth analysis of the fundamentals of a smaller number of securities.
  - Agnostic to specific companies, quantitative funds trade based on patterns and dynamics across a wide swath of securities.

Investments in strategic capabilities

- Algorithmic trading strategies may further shift the investment industry from discretionary to quantitative styles
  - AQR is a quantitative investment group that relies on academic research to identify and systematically trade factors that have, over time, proven to beat the broader market.
  - AQR has begun to seek profitable patterns in markets using ML to parse through novel datasets, such as satellite pictures of shadows cast by oil wells and tankers.

Chapter 2 Market and Fundamental Data-Sources and Techniques

Market and Fundamental Data – Sources and Techniques

- How market data reflects the structure of the trading environment
- Working with trade and quote data at minute frequency
- Reconstructing an order book from tick data using Nasdaq ITCH
- Summarizing tick data using various types of bars
- Working with **eXtensible Business Reporting Language(XBRL)**-encoded electronic filings
- Parsing and combining market and fundamental data to create a **price-to-earnings(P/E)** series
- How to access various market and fundamental data sources using Python

Market data reflects its environment

- Market data
  - How orders are processed
  - How prices are set by matching demand and supply
- The data reflects the institutional environment of trading venues
  - Rules and regulations
  - Trade execution
  - Price formation
- Algorithmic traders use algorithms to analyze volume and price

Market microstructure

- How to trade
  - Market order
  - Limit order
  - Stop order
- Where to trade
  - Exchange
  - Dark pool

Working with high-frequency data

- Consolidated feed
- Proprietary products
- Nasdaq order book data
  - **Level 1:** Real-time price information, as available from numerous online sources
  - **Level 2:** Adds information about bid and ask prices by specific market makers
  - **Level 3:** Adds the ability to enter or change quotes, execute orders, and confirm trades

Remote data access using pandas

- Download html table
- pandas-datareader for Market Data
  - Yahoo Finance
  - Investor Exchange (IEX)
  - Book Data
  - Quandl
  - FRED
  - Fama/French
  - World Bank
  - OECD
  - NASDAQ Symbols
  - Tiingo

Remote data access using pandas

- API sources:
  - Tiingo: Historical end-of-day prices on equities, mutual funds, and ETF.
  - IEX: Historical stock prices are available if traded on IEX.
  - Alpha Vantage: Historical equity data for daily, weekly, and monthly frequencies, 20+ years, and the past 3-5 days of intraday data. It also has FOREX and sector performance data.
  - Quandl: Free data sources as listed on their website.
  - Fama/French: Risk factor portfolio returns.
  - TSP Fund Data: Mutual fund prices.
  - Nasdaq: Latest metadata on traded tickers.
  - Stooq Index Data: Some equity indices are not available from elsewhere due to licensing issues.
  - MOEX: Moscow Exchange historical data.

Zipline

- It is also available offline to develop a strategy using a limited number of free data bundles that can be ingested and used to test the performance of trading ideas before porting the result to the online Quantopian platform for paper and live trading.

Fundamental Data

- EDGAR(Electronic Data Gathering, Analysis, and Retrieval) system
- Case: Building a price/earnings time series
  - Retrieving all quarterly Apple filings
  - Adjusting the earnings data to get the P/E series

Efficient data storage with pandas

- CSV: Comma-separated, standard flat text file format.
- HDF5: Hierarchical data format, developed initially at the National Center for Supercomputing Applications. It is a fast and scalable storage format for numerical data, available in pandas using the PyTables library.
- Parquet: Part of the Apache Hadoop ecosystem, a binary, columnar storage format that provides efficient data compression and encoding and has been developed by Cloudera and Twitter. It is available for pandas through the pyarrow library, led by Wes McKinney, the original author of pandas.
- purely numerical data: HDF5
- mix of numerical and text data: Parquet