Backtesting II

Marcos López de Prado, Ph.D.
Advances in Financial Machine Learning
ORIE 5256
What are we going to learn today?

• Backtest Statistics
  o General Characteristics
  o Performance
  o Time-Weighted Rate of Return
  o Drawdown and Time Under Water
  o Implementation Shortfall
  o Efficiency
  o Classification Scores

• Understanding Strategy Risk
  o Symmetric Payouts
  o Asymmetric Payouts
  o The Probability of Strategy Failure
Backtest Statistics
General Statistics

- **Time range**: Time range specifies the start and end dates.
- **Average AUM**: This is the average dollar value of the assets under management.
- **Capacity**: A strategy’s capacity can be measured as the highest AUM that delivers a target risk-adjusted performance.
- **Leverage**: Leverage measures the amount of borrowing needed to achieve the reported performance.
- **Maximum dollar position size**: Maximum dollar position size informs us whether the strategy at times took dollar positions that greatly exceeded the average AUM.
- **Ratio of longs**: The ratio of longs show what proportion of the bets involved long positions.
- **Frequency of bets**: The frequency of bets is the number of bets per year in the backtest.
- **Average holding period**: The average holding period is the average number of days a bet is held.
- **Annualized turnover**: Annualized turnover measures the ratio of the average dollar amount traded per year to the average annual AUM.
- **Correlation to underlying**: This is the correlation between strategy returns and the returns of the underlying investment universe.
Performance

- **PnL**: The total amount of dollars (or the equivalent in the currency of denomination) generated over the entirety of the backtest, including liquidation costs from the terminal position.

- **PnL from long positions**: The portion of the PnL dollars that was generated exclusively by long positions.

- **Annualized rate of return**: The time-weighted average annual rate of total return, including dividends, coupons, costs, etc.

- **Hit ratio**: The fraction of bets that resulted in a positive PnL.

- **Average return from hits**: The average return from bets that generated a profit.

- **Average return from misses**: The average return from bets that generated a loss.
Time-Weighted Rate of Return (1/2)

- The TWRR for portfolio $i$ between subperiods $[t - 1, t]$ is denoted $r_{i,t}$, with equations

$$r_{i,t} = \frac{\pi_{i,t}}{K_{i,t}}; \quad \pi_{i,t} = \sum_{j=1}^{J} [(\Delta P_{j,t} + A_{j,t})\theta_{i,j,t-1} + \Delta\theta_{i,j,t}(P_{j,t} - \bar{P}_{j,t-1})]$$

$$K_{i,t} = \sum_{j=1}^{J} \bar{P}_{j,t-1}\theta_{i,j,t-1} + \max \left\{ 0, \sum_{j=1}^{J} \bar{P}_{j,t}\Delta\theta_{i,j,t} \right\}$$

where

- $\pi_{i,t}$ is the mark-to-market (MtM) profit or loss for portfolio $i$ at time $t$.
- $K_{i,t}$ is the market value of the assets under management by portfolio $i$ through subperiod $t$. The purpose of including the max{} term is to fund additional purchases (ramp-up).
- $A_{j,t}$ is the interest accrued or dividend paid by one unit of instrument $j$ at time $t$.
- $P_{j,t}$ is the clean price of security $j$ at time $t$.
- $\theta_{i,j,t}$ are the holdings of portfolio $i$ on security $j$ at time $t$. 
Time-Weighted Rate of Return (2/2)

... where (continued)

- $\bar{P}_{j,t}$ is the dirty price of security $j$ at time $t$.
- $\bar{P}_{j,t}$ is the average transacted clean price of portfolio $i$ on security $j$ over subperiod $t$.
- $\bar{P}_{j,t}$ is the average transacted dirty price of portfolio $i$ on security $j$ over subperiod $t$.

- Inflows are assumed to occur at the beginning of the day, and outflows are assumed to occur at the end of the day. These sub-period returns are then linked geometrically as

$$\varphi_{i,T} = \prod_{t=1}^{T} (1 + r_{i,t})$$

- The variable $\varphi_{i,T}$ can be understood as the performance of one dollar invested in portfolio $i$ over its entire life, $t = 1, \ldots, T$.

Finally, the annualized rate of return of portfolio $i$ is

$$R_i = \left(\varphi_{i,T}\right)^{-\frac{1}{y_i}} - 1$$

where $y_i$ is the number of years elapsed between $r_{i,1}$ and $r_{i,T}$. 
Drawdown and Time Under Water

• Intuitively, a **drawdown** (DD) is the maximum loss suffered by an investment between two consecutive high-watermarks (HWMs).

• The **time under water** (TuW) is the time elapsed between an HWM and the moment the PnL exceeds the previous maximum PnL.

```python
def computeDD_TuW(series,dollars=False):
    # compute series of drawdowns and the time under water associated with them
    df0=series.to_frame('pnl')
    df0['hwm']=series.expanding().max()
    df1=df0.groupby('hwm').min().reset_index()
    df1.columns=['hwm','min']
    df1.index=df0['hwm'].drop_duplicates(keep='first').index
    # time of hwm
    df1[df1['hwm']>df1['min']] # hwm followed by a drawdown
    if dollars:dd=df1['hwm']-df1['min']
    else:dd=1-df1['min']/df1['hwm']
    tuw=((df1.index[1:]-df1.index[:-1])/np.timedelta64(1,'Y')).values # in years
    tuw=pd.Series(tuw,index=df1.index[:-1])
    return dd,tuw
```

![Diagram showing drawdown and time under water](image_url)
Implementation Shortfall

• **Broker fees per turnover**: These are the fees paid to the broker for turning the portfolio over, including exchange fees.

• **Average slippage per turnover**: These are execution costs, excluding broker fees, involved in one portfolio turnover.

• **Dollar performance per turnover**: This is the ratio between dollar performance (including brokerage fees and slippage costs) and total portfolio turnovers.

• **Return on execution costs**: This is the ratio between dollar performance (including brokerage fees and slippage costs) and total execution costs.
Efficiency

• **Annualized Sharpe ratio**: This is the SR value, annualized by a factor $\sqrt{a}$, where $a$ is the average number of returns observed per year.

• **Information ratio**: This is the SR equivalent of a portfolio that measures its performance relative to a benchmark.

• **Probabilistic Sharpe ratio**: PSR corrects SR for inflationary effects caused by non-Normal returns or track record length.

• **Deflated Sharpe ratio**: DSR corrects SR for inflationary effects caused by non-Normal returns, track record length, and selection bias under multiple testing.
Sharpe [1966]

- Consider an investment strategy with excess returns (or risk premia) \( \{r_t\}, t = 1, \ldots, T \), which follow an IID Normal distribution,

\[
r_t \sim \mathcal{N}[\mu, \sigma^2]
\]

where \( \mathcal{N}[\mu, \sigma^2] \) represents a Normal distribution with mean \( \mu \) and variance \( \sigma^2 \).
- The SR (non-annualized) of such strategy is defined as

\[
SR = \frac{\mu}{\sigma}
\]

- Because parameters \( \mu \) and \( \sigma \) are not known, SR is estimated as

\[
\hat{SR} = \frac{E[\{r_t\}]}{\sqrt{V[\{r_t\}]}}
\]
Lo [2002]

• Under the assumption that returns follow an IID Normal distribution, Lo [2002] derived the asymptotic distribution of $\hat{SR}$ as

$$\left( \hat{SR} - SR \right)^{\alpha} \to N \left[ 0, \frac{1 + \frac{1}{2} SR^2}{T} \right]$$

• Under the assumption that returns follow an IID non-Normal distribution, Mertens [2002] derived the asymptotic distribution of $\hat{SR}$ as

$$\left( \hat{SR} - SR \right)^{\alpha} \to N \left[ 0, \frac{1 + \frac{1}{2} SR^2 - \gamma_3 SR + \frac{\gamma_4 - 3}{4} SR^2}{T} \right]$$

where $\gamma_3$ is the skewness of $\{r_t\}$, and $\gamma_4$ is the kurtosis of $\{r_t\}$ ($\gamma_3 = 0$ and $\gamma_4 = 3$ when returns follow a Normal distribution).
Bailey and López de Prado [2012] (1/2)

• Christie [2005] and Opdyke [2007] discovered that, in fact, the Mertens [2002] equation is also valid under the more general assumption that returns are stationary and ergodic (not necessarily IID).

• Bailey and López de Prado [2012] utilized those results to derive the **Probabilistic Sharpe Ratio** (PSR).

• PSR estimates the probability that an observed $\hat{SR}$ exceeds $SR^*$ as

$$PSR[SR^*] = Z \left[ \frac{\hat{SR} - SR^*}{\sqrt{1 - \hat{\gamma}_3 \hat{SR} + \frac{\hat{\gamma}_4 - 1}{4} \hat{SR}^2}} \right]$$

where $Z[\cdot]$ is the CDF of the standard Normal distribution, $T$ is the number of observed returns, $\hat{\gamma}_3$ is the skewness of the returns, and $\hat{\gamma}_4$ is the kurtosis of the returns. Note that $\hat{SR}$ is the non-annualized estimate of SR, computed on the same frequency as the $T$ observations.
• For a given $SR^*$, $\hat{PSR}$ increases with
  • greater mean returns ($E\{r_t\}$)
  • lower variance of returns ($V\{r_t\}$)
  • longer track records ($T$)
  • positively skewed returns ($\hat{\gamma}_3$)
  • thinner tails ($\hat{\gamma}_4$)

• This result also allows us to answer the question: “How long should a track record be in order to have statistical confidence $(1 - \alpha)$ that its estimated Sharpe ratio ($\hat{SR}$) is above a given threshold ($SR^*$)” (minimum track record length)

$$MinTRL = 1 + \left[1 - \hat{\gamma}_3\hat{SR} + \frac{\hat{\gamma}_4 - 1}{4} \hat{SR}^2\right] \left(\frac{Z_\alpha}{\hat{SR} - SR^*}\right)^2$$

where $Z_\alpha$ is the value of the Standard Normal CDF that leaves a probability $\alpha$ in the right tail.
Bailey and López de Prado [2014] (1/2)

- **The Deflated Sharpe Ratio** computes the probability that the Sharpe Ratio (SR) is statistically significant, after controlling for the inflationary effect of multiple trials, data dredging, non-normal returns and shorter sample lengths.

\[
DSR = P_{SR} \left( \tilde{SR}_0 \right) = Z \left[ \frac{(SR - \tilde{SR}_0)\sqrt{T} - 1}{\sqrt{1 - \hat{\gamma}_3 SR + \frac{\gamma^4}{4} SR^2}} \right]
\]

where \( \tilde{SR}_0 \) is the estimate provided by the False Strategy theorem,

\[
\tilde{SR}_0 = \sqrt{V\left[\left\{ \overline{SR}_k \right\} \right]} \left( (1 - \gamma)Z^{-1} \left[ 1 - \frac{1}{K} \right] + \gamma Z^{-1} \left[ 1 - \frac{1}{Ke} \right] \right)
\]

- DSR packs more information than SR, and it is expressed in probabilistic terms.
Bailey and López de Prado [2014] (2/2)

• The standard SR is computed as a function of two estimates:
  • Mean of returns
  • Standard deviation of returns

• DSR deflates SR by taking into consideration five additional variables (it packs more information):
  • The non-Normality of the returns ($\gamma_3, \gamma_4$)
  • The length of the returns series ($T$)
  • The amount of data dredging ($V[\{SR_k\}]$)
  • The number of independent trials involved in the selection of the investment strategy ($K$)

The key to preventing selection bias is to record all trials, and determine correctly the number of effectively independent trials ($K$).
Classification Scores

- **Accuracy**: The fraction of opportunities correctly labeled.
- **Precision**: The fraction of true positives among the predicted positives.
- **Recall**: The fraction of true positives among the positives.
- **F1**: The (equally weighted) harmonic mean of precision and recall.
- **Log-loss (cross-entropy loss)**: It computes the log-likelihood of the classifier given the true label, which takes predictions’ probabilities into account. Log loss can be estimated as follows:

\[
L[Y, P] = -\log[\text{Prob}[Y|P]] = -N^{-1} \sum_{n=0}^{N-1} \sum_{k=0}^{K-1} y_{n,k} \log[p_{n,k}]
\]

where:
- \(p_{n,k}\) is the probability associated with prediction \(n\) of label \(k\).
- \(Y\) is a 1-of-\(K\) binary indicator matrix, such that \(y_{n,k} = 1\) when observation \(n\) was assigned label \(k\) out of \(K\) possible labels, and 0 otherwise.
Understanding Strategy Risk
Symmetric Payouts (1/2)

• Consider a strategy that produces \( n \) IID bets per year, where the outcome \( X_i \) of a bet \( i \in [1, n] \) is a profit \( \pi > 0 \) with probability \( P[X_i = \pi] = p \), and a loss \(-\pi\) with probability \( P[X_i = -\pi] = 1 - p \).

• Think of \( p \) as the precision of a binary classifier where a positive means betting on an opportunity, and a negative means passing on an opportunity: True positives are rewarded, false positives are punished, and negatives (whether true or false) have no payout.

• Since the betting outcomes \( \{X_i\}_{i=1}^n \) are independent, we compute the expected moments per bet:
  
  o The expected profit from one bet is \( E[X_i] = \pi p + (-\pi)(1 - p) = \pi (2p - 1) \).
  
  o The variance is \( V[X_i] = E[X_i^2] - E[X_i]^2 \), where \( E[X_i^2] = \pi^2 p + (-\pi)^2 (1 - p) = \pi^2 \), thus \( V[X_i] = \pi^2 - \pi^2 (2p - 1)^2 = \pi^2 [1 - (2p - 1)^2] = 4\pi^2 p(1 - p) \).

• For \( n \) IID bets per year, the annualized Sharpe ratio (\( \theta \)) is

\[
\theta[p, n] = \frac{nE[X_i]}{\sqrt{nV[X_i]}} = \frac{2p - 1}{2\sqrt{p(1 - p)}} \sqrt{n}
\]

\[t\text{-value of } p \text{ under } H_0: p = \frac{1}{2}\]
Symmetric Payouts (2/2)

• Note how $\pi$ cancels out of the above equation, because the payouts are symmetric.

• Just as in the Gaussian case, $\theta[p, n]$ can be understood as a re-scaled $t$-value.

• This illustrates the point that, even for a small $p > \frac{1}{2}$, the Sharpe ratio can be made high for a sufficiently large $n$.

• This is the economic basis for high-frequency trading, where $p$ can be barely above .5, and the key to a successful business is to increase $n$.

• The Sharpe ratio is a function of precision rather than accuracy, because passing on an opportunity (a negative) is not rewarded or punished directly (although too many negatives may lead to a small $n$, which will depress the Sharpe ratio toward zero).

  o For example, $p = .55 \implies \frac{2p-1}{2\sqrt{p(1-p)}} = 0.1005$, and achieving an annualized Sharpe ratio of 2 requires 396 bets per year.
Asymmetric Payouts

• Consider a strategy that produces $n$ IID bets per year, where the outcome $X_i$ of a bet $i \in [1, n]$ is $\pi_+$ with probability $P[X_i = \pi_+] = p$, and an outcome $\pi_-$, $\pi_- < \pi_+$ occurs with probability $P[X_i = \pi_-] = 1 - p$.
  
  o The expected profit from one bet is $E[X_i] = p\pi_+ + (1 - p)\pi_- = (\pi_+ - \pi_-)p + \pi_-$.  
  
  o The variance is $V[X_i] = E[X_i^2] - E[X_i]^2$, where $E[X_i^2] = p\pi_+^2 + (1 - p)\pi_-^2 = (\pi_+^2 - \pi_-^2)p + \pi_-^2$, thus $V[X_i] = (\pi_+ - \pi_-)^2p(1 - p)$.

• For $n$ IID bets per year, the annualized Sharpe ratio ($\theta$) is

$$
\theta[p, n, \pi_-, \pi_+] = \frac{nE[X_i]}{\sqrt{nV[X_i]}} = \frac{(\pi_+ - \pi_-)p + \pi_-}{(\pi_+ - \pi_-)\sqrt{p(1 - p)}} \sqrt{n}
$$

and for $\pi_- = -\pi_+$ we can see that this equation reduces to the symmetric case:

$$
\theta[p, n, -\pi_+, \pi_+] = \frac{2\pi_+ p + \pi_+}{2\pi_+ \sqrt{p(1-p)}} \sqrt{n} = \frac{2p - 1}{2\sqrt{p(1-p)}} \sqrt{n} = \theta[p, n].
$$

• For example, for $n = 260$, $\pi_- = -.01$, $\pi_+ = .005$, $p = .7$, we get $\theta = 1.173$.  

The Probability of Strategy Failure (1/2)

• In the example above, parameters
  o \( \pi_- = -0.01, \pi_+ = 0.005 \) are set by the portfolio manager, and passed to the traders with the execution orders.
  o Parameter \( n = 260 \) is also set by the portfolio manager, as she decides what constitutes an opportunity worth betting on.

• The two parameters that are not under the control of the portfolio manager are \( p \) (determined by the market) and \( \theta^* \) (the objective set by the investor). Because \( p \) is unknown, we can model it as a random variable, with expected value \( E[p] \).

• Let us define \( p_{\theta^*} \) as the value of \( p \) below which the strategy will underperform a target Sharpe ratio \( \theta^* \), that is, \( p_{\theta^*} = \max\{p|\theta \leq \theta^*\} \).

• For \( p_{\theta^*=0} = \frac{2}{3}, p < p_{\theta^*=0} \Rightarrow \theta \leq 0 \). This highlights the risks involved in this strategy, because a relatively small drop in \( p \) (from \( p = 0.7 \) to \( p = 0.67 \)) will wipe out all the profits. The strategy is intrinsically risky, even if the holdings are not.

• That is a critical difference missing in most asset management textbooks: Strategy risk should not be confused with portfolio risk.
The Probability of Strategy Failure (2/2)

• Firms and investors compute, monitor, and report portfolio risk without realizing that this tells us nothing about the risk of the strategy itself.

• Strategy risk is not the risk of the underlying portfolio, as computed by the chief risk officer.

• Strategy risk is the risk that the strategy will fail to succeed over time, a question of far greater relevance to the chief investment officer.

• The answer to the question “What is the probability that this strategy will fail?” is equivalent to computing $P[p < p_{\theta^*}]$. 
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.

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.

THANKS FOR YOUR ATTENTION!
Bio

Dr. Marcos López de Prado is a principal at AQR Capital Management, and its head of machine learning. Before AQR, he founded and led Guggenheim Partners’ Quantitative Investment Strategies (QIS) business, where he developed high-capacity machine learning strategies that consistently delivered superior risk-adjusted returns, receiving up to $13 billion in assets.

Concurrently with the management of investments, between 2011 and 2018 Marcos was also a research fellow at Lawrence Berkeley National Laboratory (U.S. Department of Energy, Office of Science). He has published dozens of scientific articles on machine learning and supercomputing in the leading academic journals, and SSRN ranks him as one of the most-read authors in economics. Among several monographs, he is the author of the graduate textbook Advances in Financial Machine Learning (Wiley, 2018).

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 post-doctoral research at Harvard University and Cornell University, where he teaches a financial machine learning course at the School of Engineering. Marcos has an Erdős #2 and an Einstein #4 according to the American Mathematical Society.

For more information, please visit www.QuantResearch.org
Disclaimer

• The views expressed in this document are the authors’ and do not necessarily reflect those of the organizations he is affiliated with.

• No investment decision or particular course of action is recommended by this presentation.

• All Rights Reserved. © 2018 by Marcos López de Prado