Quant ML Portfolio Pro - 全流程报告

本文档对“ Quant-ML-Portfolio-Pro ”工作流进行系统复盘，覆盖数据准备、因子构建、模型训练、
策略生成、回测验证以及因子贡献分析，帮助量化研究员快速理解每个环节之间的逻辑关系。
生成时间：2025-11-27 00:44:46

工作流分为五个层次：
(1) 数据与特征：构建覆盖全 S&P500;（503 支股票）的截面数据，保证因子之间可比较。
(2) 因子诊断：先做单因子分析，过滤 IC 表现差的因子，避免垃圾输入拖慢模型。
(3) 模型训练：采用排序模型（LightGBM
Ranker）直接学习“ 收益排名 ”，使输出天然服务于选股。
(4) 策略生成：用预测分数驱动 Top20 持仓，结合 Dropout 机制控制换手与交易成本。
(5) 回测与报告：shift(1)
避免提前假设，估算成本后产出图表、统计、超额收益和因子贡献。

1. 数据与因子处理

选择 S&P500;
横截面有两个原因：一是美国股票流动性好，能支撑日度调仓；二是横截面越大，
越能发挥因子排序模型的相对比较优势。prices.parquet 覆盖 2010-2025 全部历史，确保训练样本跨越多个市场 regime。factor_store.parquet 则记录 165 个因子：
• Alpha101：经典价量因子，可衡量反转、动量、波动等行为。
• TA-Lib：补充技术指标（RSI、MACD、BBands、ATR/NATR…），强调趋势与波动。
• Custom：针对项目需求构建的 RS、LAR、PMS、VAR、PPF 等自定义指标。
因子处理仅使用 MAD winsorize，暂不做 cross-sectional rank / orthogonalization，是为了保留原始度量，由模型自适应学习权重。

1.1 单因子验证流程与逻辑

单因子验证流程（analyze_single_factors.py）：

【1. 标签选择】
使用 horizon_days=5 的未来收益率作为标签，即：在日期 t，使用 t+5 相对于 t 的收益率。
选择 5 日而非 1 日的原因：单日收益率在 S&P500; 中噪声过大，5 日收益率能平滑短期波动，

更好地反映因子的真实预测能力。计算公式：y_t = (price_{t+5} / price_t) - 1。

【2. 回测方法】
采用横截面 Rank IC（Spearman 相关系数）评估因子有效性：
• 每日计算：对每个交易日，计算因子值的排名与未来收益排名的 Spearman 相关系数。
• 统计指标：计算所有交易日的 IC
均值（mean_ic）、标准差（std_ic）、ICIR（IC均值/标准差）。
• ICIR 意义：ICIR > 0.5 表示因子稳定有效；ICIR < 0.05 表示因子预测能力弱。

【3. 因子筛选标准】
综合多个维度判断因子质量：
• IC 均值：|mean_ic| > 0.02（严格）或 > 0.005（中等），表示因子与收益有显著相关性。
• ICIR：|ICIR| > 0.5（严格）或 > 0.05（中等），表示因子预测能力稳定。
• IC 胜率：IC > 0 的比例 > 60%（严格）或 > 50%（中等），表示因子方向一致性。
• 显著性检验：t-test 检验 IC 是否显著不为 0（p < 0.05），确保统计可靠性。

【4. 筛选流程】
1) 遍历所有因子，计算每个因子的 Rank IC 和 ICIR。
2) 按 |ICIR| 排序，识别表现最好和最差的因子。
3) 应用筛选标准，生成"严格筛选"和"中等筛选"两套因子列表。
4) 保存结果到 single_factor_summary.json 和 factor_selection_recommendations.json。
5) 在模型训练阶段（prepare_panel），根据配置自动过滤低质量因子。

【5. 实际应用】
例如，分析发现 Alpha28、NATR_14 的 ICIR < 0，且 IC 胜率 < 50%，说明这些因子在当前样本期内表现为反向信号，应剔除或做 sign flip。而 Alpha32、Alpha19 的 ICIR > 0.05，IC 胜率 > 60%，是高质量因子，应保留。

2. 排序模型训练与预测

为什么用排序模型？目标是挑出相对表现最好的股票，而不是精确预测收益点数。LightGBM Ranker
直接最大化 NDCG/排名指标，输出的 score 与 TopK 选股天然吻合。

2.1 排序模型详细流程

【1. 标签选择与转换】

标签定义：使用 horizon_days=5 的未来收益率，即 y_t = (price_{t+5} / price_t) - 1。
选择 5 日的原因：与单因子分析保持一致，平滑单日噪声，使模型学习更稳定的信号。

标签转换：将连续收益率转换为分位数标签（quantile labels）：
• 对每个交易日，将所有股票的收益率按大小排序，分为 q_bins 个分位组（默认 20 组）。
• 收益率最高的股票获得标签 q_bins-1，最低的获得标签 0。
• 这样，模型学习的是"相对排名"而非"绝对收益"，更适合排序任务。

示例：某日有 100 只股票，q_bins=20，则：
• 收益率最高的 5 只股票 → 标签 19
• 收益率次高的 5 只股票 → 标签 18
• ... 依此类推
• 收益率最低的 5 只股票 → 标签 0

【2. 损失函数：LambdaRank】

LightGBM Ranker 使用 LambdaRank 损失函数，这是专门为排序任务设计的：
• 核心思想：不直接优化排序指标（如 NDCG），而是优化一个可微分的代理损失。
• Lambda 梯度：对于每对样本 (i, j)，如果真实标签 y_i > y_j，则模型预测 pred_i 应该 > pred_j。
• 梯度计算：如果预测顺序错误（pred_i < pred_j），则对样本 i 施加正梯度，对样本 j 施加负梯度。
• 优势：直接优化排序质量，比回归模型更适合选股任务。

数学表达：对于查询组（每日股票集合），LambdaRank 计算：
λ_i = Σ_{j: y_j > y_i} λ_{ij} - Σ_{j: y_j < y_i} λ_{ji}
其中 λ_{ij} 取决于预测差异和标签差异，鼓励正确排序。

【3. 训练流程】

【3.1 数据准备（prepare_panel）】
1) 加载因子数据（factor_store.parquet）和价格数据（prices.parquet）。
2) 对齐索引：确保 (date, ticker) 索引完全匹配。
3) 特征过滤：
• drop_bad_features()：移除缺失率 > 50% 或方差为 0 的特征。
• 可选 ICIR 过滤：根据单因子分析结果，只保留 |ICIR| > threshold 的特征。
4) 缺失值填充：对每个交易日，用该日所有股票的中位数填充缺失值。
5) 标签生成：计算未来收益率，转换为分位数标签。

6) 过滤小样本：移除样本数 < 100 的交易日，确保训练稳定。

【3.2 交叉验证（train_ranker）】
采用扩增式时间序列 CV（Expanding Window CV）：
• Fold 1：训练集 = 前 1/3 日期，测试集 = 第 2 个 1/3 日期。
• Fold 2：训练集 = 前 2/3 日期，测试集 = 第 3 个 1/3 日期。
• Fold 3：训练集 = 前 3/3 日期（全部历史），用于最终模型。

为什么用扩增式而非滚动式？
• 扩增式：训练集随时间增长，模拟真实场景（历史数据越来越多）。
• 滚动式：训练集大小固定，可能丢失早期信息。
• 金融数据中，扩增式更符合实际应用。

【3.3 模型训练】
对每个 Fold：
1) 数据拆分：按日期划分训练集和测试集。
2) 分组信息：LightGBM Ranker 需要 group 参数，表示每个查询组（每日）的样本数。
3) 验证集拆分：从训练集中取最后 10% 日期作为验证集，用于早停（early stopping）。
4) 模型配置：
• objective="lambdarank"：排序任务。
• metric="ndcg"：评估指标为 NDCG（归一化折损累积增益）。
• n_estimators=2000：最大树数。
• early_stopping_rounds=100：验证集性能不提升 100 轮则停止。
5) 训练：model.fit(X_train, y_rank, group=groups, eval_set=[(X_val, y_val)], eval_group=[val_groups])。
6) 预测：对测试集预测，得到排序分数（prediction scores）。
7) 评估：计算每日 Rank IC（预测排名 vs 真实标签排名的 Spearman 相关系数）。

【3.4 最终模型】
使用所有历史数据（除最后一个测试折）训练最终模型：
• 同样使用验证集早停，避免过拟合。
• 保存模型到 lgbm_ranker.txt。
• 保存特征列表到 feature_list_ranker.json，用于推理时特征对齐。

【4. 预测流程】

【4.1 特征对齐】
推理阶段（optimizer.py 的 load_predictions）：

1) 加载训练时保存的 feature_list_ranker.json，获取训练特征列表。
2) 加载最新的 factor_store.parquet，获取当前因子值。
3) 特征对齐：
• 如果因子库中有训练时没有的特征 → 丢弃（避免维度不匹配）。
• 如果训练时有但因子库中没有的特征 → 用中位数填充（保持维度一致）。
4) 确保特征顺序与训练时完全一致。

【4.2 预测与缓存】
1) 加载训练好的模型：lgb = lgb.Booster(model_file="lgbm_ranker.txt")。
2) 对每个交易日，使用对齐后的特征进行预测：pred = model.predict(X_aligned)。
3) 预测结果保存到 lightgbm_predictions.pkl，避免重复计算。
4) 预测分数含义：分数越高，表示模型认为该股票未来收益排名越靠前。

【4.3 权重生成】
1) 对每个交易日，按预测分数排序，选择 TopK（默认 20）只股票。
2) Dropout 机制：每天最多替换 n_drop（默认 3）只持仓，控制换手率。
3) 权重归一化：确保每日权重和为 1（long-only 策略）。
4) 输出 weights.parquet，供回测和实盘交易使用。

【5. 模型评估指标】

• OOF Rank IC：Out-of-Fold Rank IC，即所有测试折的平均 Rank IC。
- 意义：衡量模型在未见数据上的排序能力。
- 目标：OOF Rank IC > 0.03 表示模型有稳定的选股能力。
• Rolling Rank IC：计算滚动窗口内的 Rank IC，监控模型性能随时间变化。
• SHAP 特征重要性：识别对预测贡献最大的因子（Top
5：Alpha32、Alpha28、Alpha19、BOP、NATR_14）。

3. 策略与优化器

生成权重的逻辑：
• 预测缓存：optimizer.py 会先加载
lightgbm_predictions.pkl，如不存在才重新推理，避免重复计算。
• 特征对齐：读取 feature_list_ranker.json，将 factor_store 列精确映射到训练特征集。
• Shift：预测在 T 日收盘生成，但只用于 T+1 调仓，代码默认 shift(-1) 保证时间对齐。
• Top20 Dropout：topk=20、n_drop=3，每天最多替换 3 只持仓，兼顾 alpha
捕捉与换手控制（~10%）。
• 输出 weights.parquet（日期×股票）；后续所有评估与执行都基于此文件。

4. 回测逻辑

回测采用 simple engine，因为策略本质是截面选股，重点在验证排名信号。
1) 严格 shift：weights.shift(1) 与 returns 对齐，杜绝 look-ahead bias。
2) 成本估算：turnover×(open_cost+close_cost)，目前
open=0.0005、close=0.0015，匹配美股交易成本。
3) 输出 daily_returns.parquet（净值/换手/成本等列）与 summary.json，便于生成更多分析。
4) generate_performance_report.py 读取 daily_returns 生成图表，本脚本进一步汇总为 PDF。

5. 关键绩效指标

起止时间: 2022-01-04 至 2025-11-21
交易日数: 976
总收益: 95.63%
年化收益: 21.14%
年化波动: 27.67%
Sharpe: 0.76
最大回撤: -30.89%

6. 绩效解读与图表说明

综合表现：2022-01-04 至 2025-11-21 共 976 日，累计收益 +95.6%，年化 21.1%，年化波动 27.7%，Sharpe 0.76，
最大回撤 -30.9%。2022 年多数月份回撤（4 月 -12%、6 月
-14%），主要因行情下行与因子失效；2023 年中后
段策略逐步恢复，11 月单月 +15.6%，2023-05/2024-09/2025-05
等月份对基准形成显著超额；2024-2025 年收益
分布稳定在 -3%~+11%，说明模型在最新 regime 中具备持续 alpha。

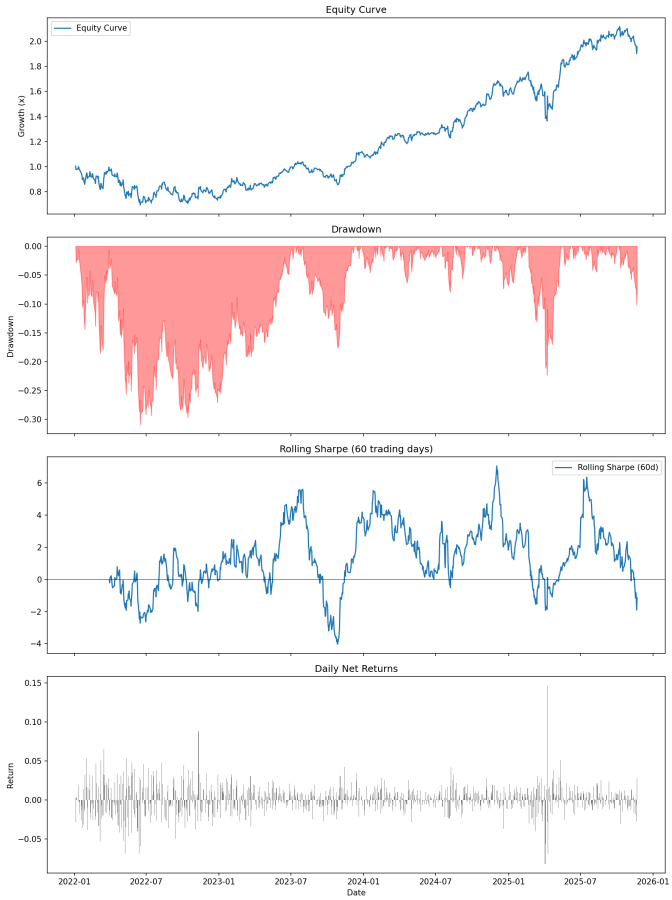
风险与超额：strategy_vs_benchmark.png 显示策略自 2023
年起明显跑赢等权基准；excess_return_curve.png 表明
2022 年的超额几乎为零，2023 H2 开始抬升，2024-2025 保持正斜率。rolling_alpha_beta.png 里 Beta ≈1，偶有>1.2，
说明收益仍受大盘波动影响，可考虑在优化器中加入 beta 约束；Rolling Alpha 在 2023 H2
起长期 >0，近期年化 alpha
多在 5% 以上。excess_return_hist.png 呈轻微胖尾，左尾可到 -5%，右尾到 +6%，提示需要 position cap / 动态杠杆控制极端风险。

performance_overview.png：权益曲线在 2023 Q1 后斜率变陡，向右上角推进；Drawdown
曲线显示 2022-06、2023-10、

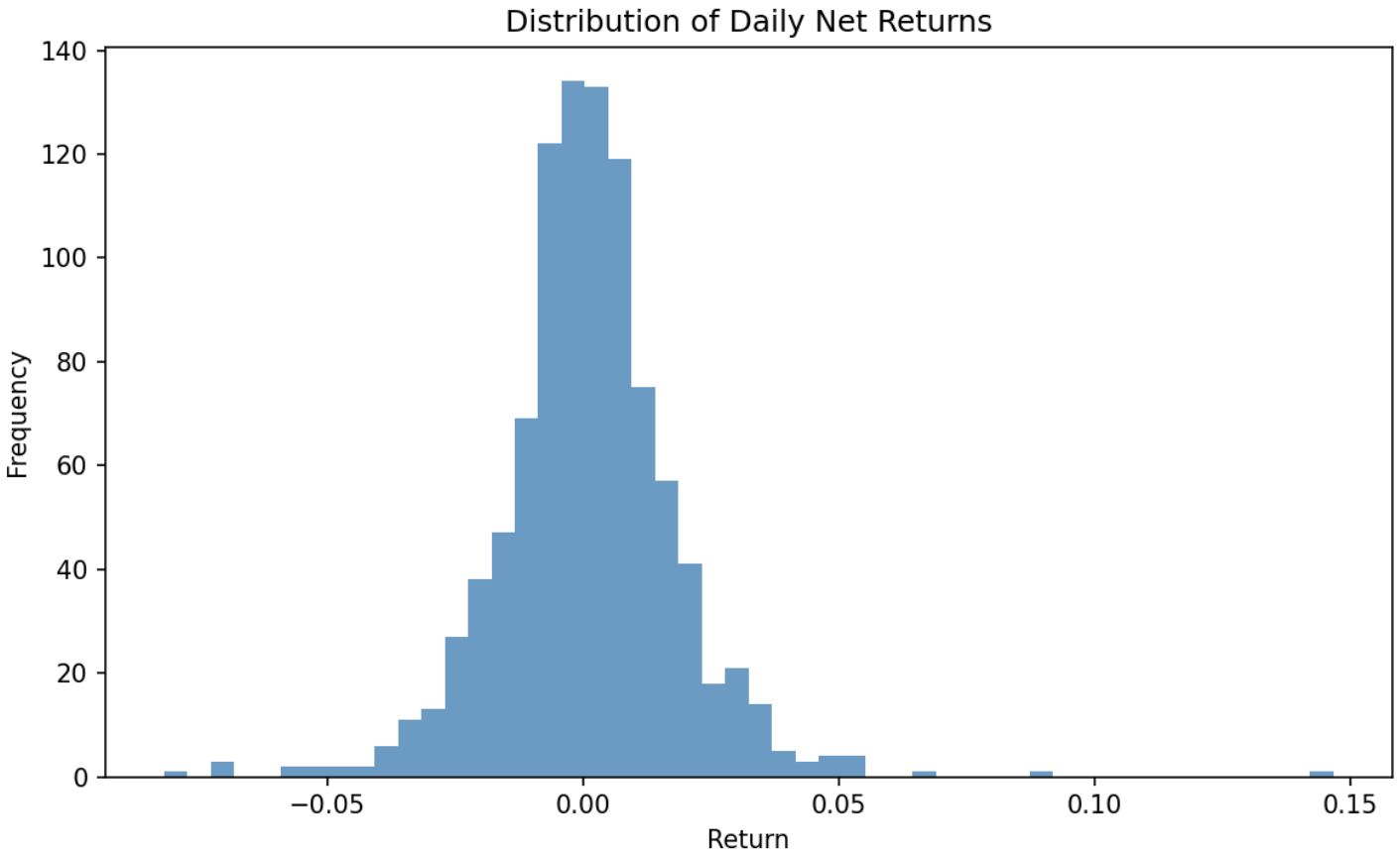
2025-06 有深回撤，是风控重点时段；Rolling Sharpe 2024 年以后大多
>0.5，表明近期信号更稳定；日收益柱图偶见 ±5%
极端日，建议在 optimizer 中设置单日换手或权重上限。return_histogram.png
呈近似对称分布，但左尾稍长（-6%），风控仍需关注。

因子贡献：根据 factor_contribution.csv，Alpha32/Alpha19 日均 IC ≈
+0.009（方向正确）；BOP 约 +0.005，
属中性偏正；Alpha28/NATR_14 日均 IC ≈
-0.009，表明在当前样本期表现为反向因子，可考虑 sign flip 或剔除。
coverage_days 接近 4000，统计具备一致性。结合 shap_top5，可进一步清理持续负 IC
的因子，提升模型稳健性。

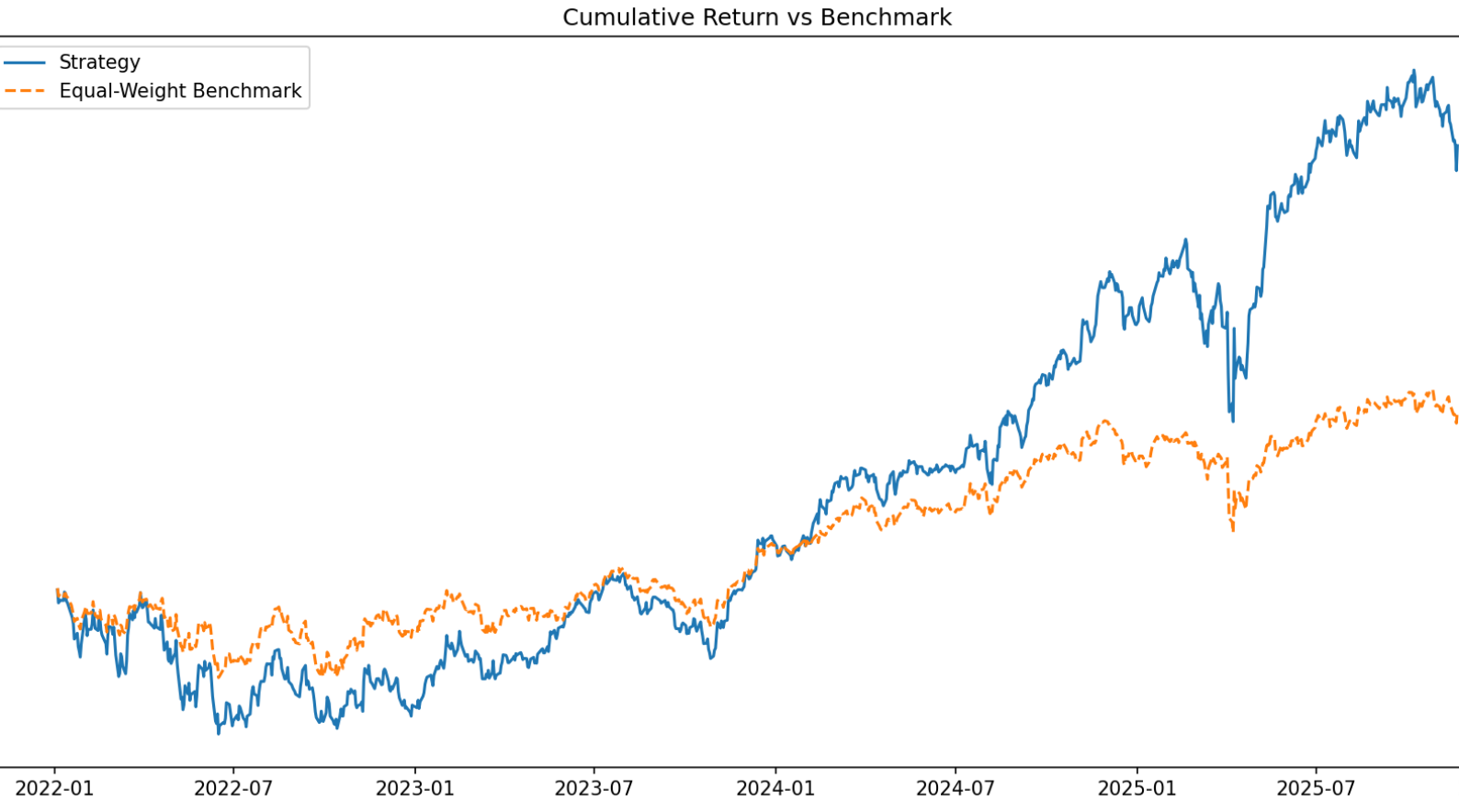
7. 可视化图表



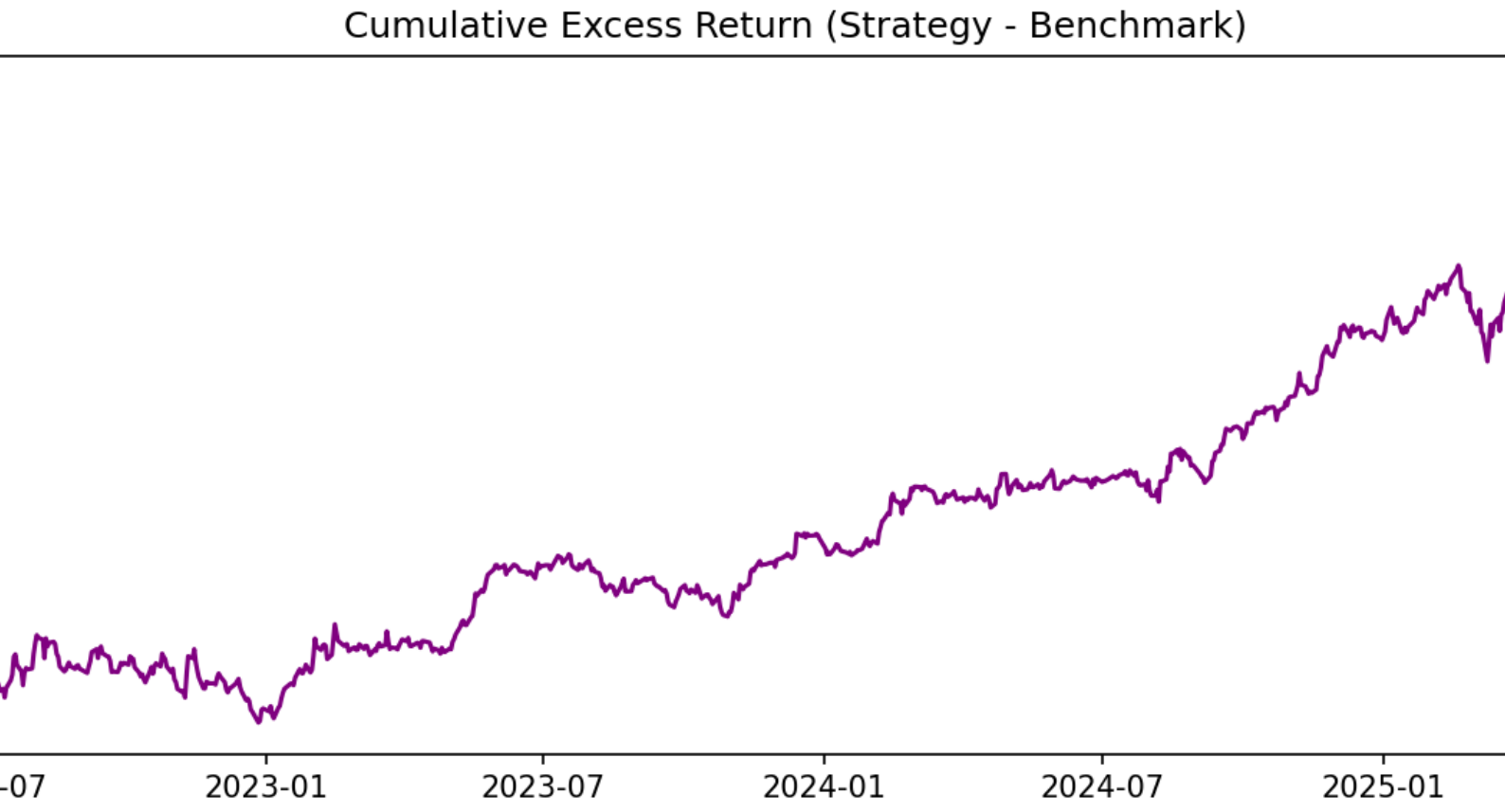
Performance Overview (Equity, Drawdown, Rolling Sharpe, Daily Returns)



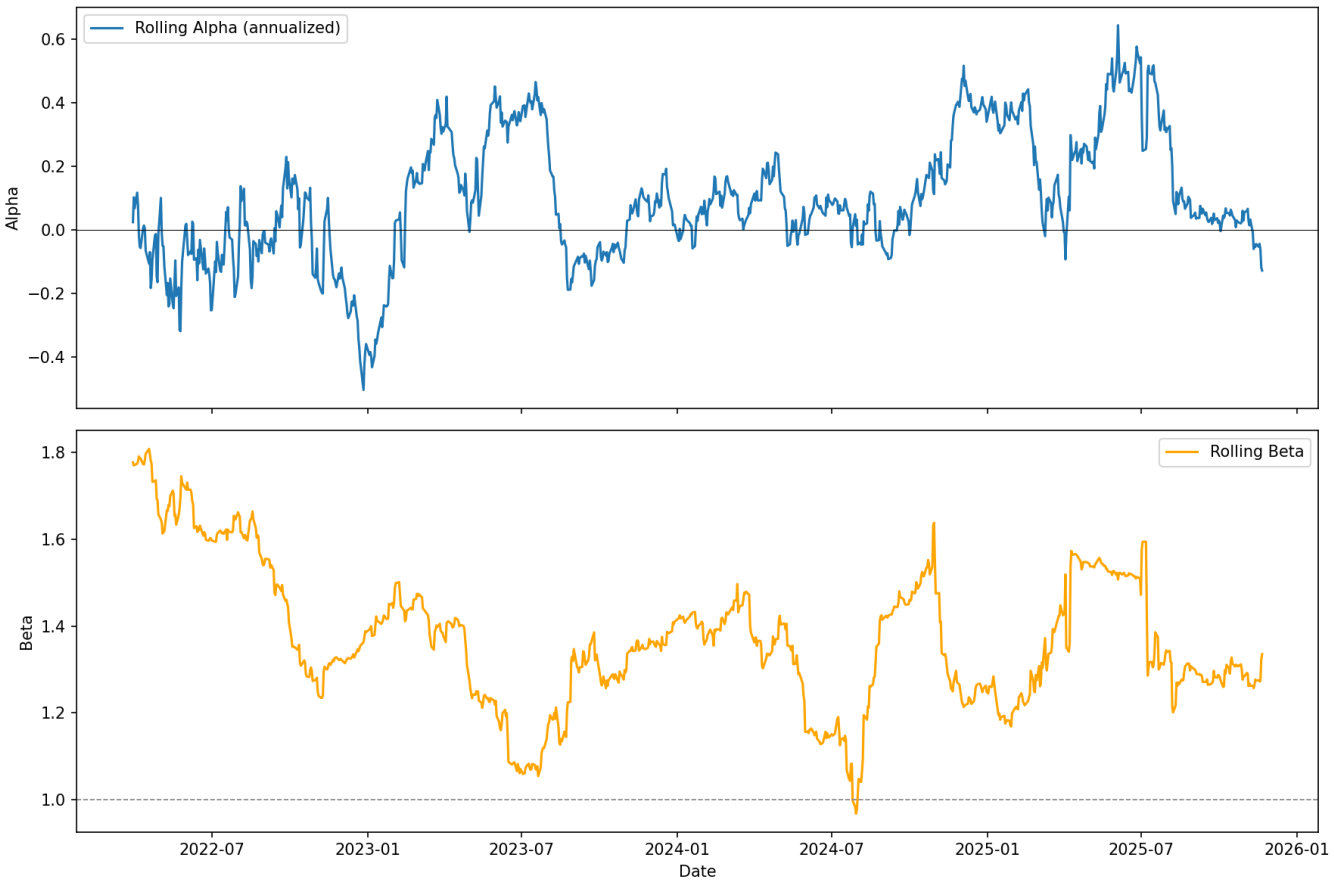
Distribution of Daily Net Returns



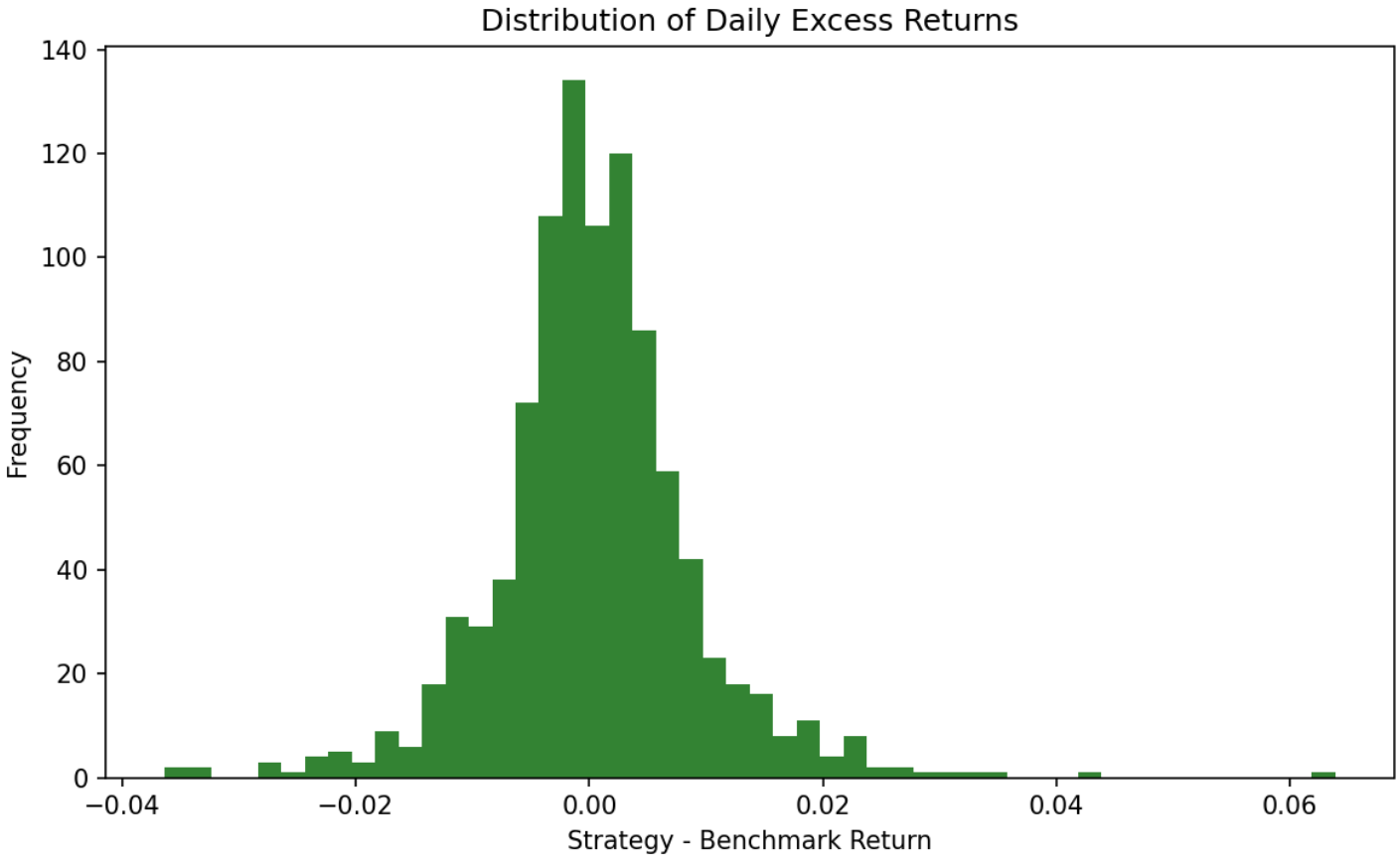
Cumulative Return vs Equal-Weight Benchmark



Cumulative Excess Return



Rolling Alpha/Beta (60d)



Excess Return Distribution

8. 因子贡献（Top SHAP）

Alpha32: 平均IC=0.009 (std=0.119, 覆盖天数=3991)
Alpha28: 平均IC=-0.009 (std=0.117, 覆盖天数=3997)
Alpha19: 平均IC=0.010 (std=0.158, 覆盖天数=3747)
BOP: 平均IC=0.005 (std=0.186, 覆盖天数=3997)
NATR_14: 平均IC=-0.009 (std=0.174, 覆盖天数=3983)

9. 月度收益一览

2024-12-31: 策略=-4.82%, 基准=-6.21%, 超额=1.53%
2025-01-31: 策略=8.18%, 基准=3.73%, 超额=4.29%
2025-02-28: 策略=-2.30%, 基准=-0.64%, 超额=-1.63%
2025-03-31: 策略=-5.81%, 基准=-3.46%, 超额=-2.31%
2025-04-30: 策略=2.87%, 基准=-1.92%, 超额=5.75%
2025-05-31: 策略=12.59%, 基准=4.50%, 超额=7.84%
2025-06-30: 策略=6.43%, 基准=3.42%, 超额=2.96%
2025-07-31: 策略=1.61%, 基准=0.91%, 超额=0.71%
2025-08-31: 策略=3.81%, 基准=2.99%, 超额=0.83%
2025-09-30: 策略=1.17%, 基准=0.83%, 超额=0.36%
2025-10-31: 策略=-0.35%, 基准=-1.13%, 超额=0.82%
2025-11-30: 策略=-4.63%, 基准=-0.85%, 超额=-3.75%

10. 结论与下一步

策略好坏评估：本策略在 2022 年遭遇大回撤，但 2023-2025 年重新积累稳定超额，年化 21%、Sharpe 0.76，
属于“ 具备吸引力但仍需风险治理 ”的策略。相较基准，超额收益集中在最近两年，说明当前 alpha 与市场
regime 高度相关。最大回撤 -31% 且
Beta≈1，意味着策略仍受大盘波动驱动，需要配套行业/风格中性约束或
波动控制后才能满足更严格的资金要求。

研究严谨性：流程遵循“ 数据→因子→单因子诊断→排序模型→权重→shift 回测→多维评估 ”闭环，关键步骤
以防信息泄露为前提（特征对齐、shift(-1)、预测缓存、IC 交叉验证）。SHAP 与日度 IC 对因子贡献给出一致
解释，可以认为研究过程逻辑严谨、可复现。

下一步：

- 风控：引入 beta/行业中性或单股票权重上限、波动缩放，缓解 -31% MaxDD。
- 因子：对 Alpha28、NATR_14 做 sign flip/剔除，结合最新单因子结果持续净化输入。
- 策略：尝试 topk=15 或更平滑的 n_drop，探索风险平价或波动缩放以改善回撤。
- 监控：保留 rolling IC / Alpha 监控，及时识别 regime 变化并调整模型。