Rethinking the ML Pipeline: Why "Train Wide, Filter Smart" is a Game-Changer for AI Factor

What if this pre-filtering is limiting our AI Factor models' potential?

ANDREAS HIMMELREICH
OCT 30, 2025

Train Wide, Filter Smart

For years, the standard approach in quantitative investing has been to pre-filter a universe (e.g., "large caps," "profitable firms") before training a machine learning model. The goal: reduce noise and focus the signal.

But what if this pre-filtering is limiting our models' potential?

The powerful alternative: Train Wide, then Filter Smart!

The Architecture Shift:

1. Train Wide: Let your (nonlinear!) model (e.g., LightGBM, ExtraTrees) learn from a broad and noisy universe. Expose it to the full market ecosystem—the quality stock the speculative ones, everything.

2. Filter Smart: Apply simple, robust quality filters (like CurFYEPSMean > 0) only a the final portfolio construction phase.

Why This Delivers Superior Results:

- Deeper Market Intelligence: A model trained on a wide universe understands regi changes, factor interactions, and market dynamics that are invisible in a pre-filtere sandbox. **It learns the context of what makes a signal valuable.**

- Transforms the Signal Distribution: This method doesn't just filter out "bad stock It fundamentally reshapes the output distribution of AI-driven factors, especially under Z-Score normalization. The result? A dramatic reduction in performance "spikiness" and significantly smoother equity curves.

- Enables Concentrated, Low-Volatility Portfolios: By cleaning the signal after the prediction, you can run highly concentrated portfolios based on ML rankings witho inheriting the volatility of the broader, noisier universe. My LightGBM strategies n exhibit the stability I previously only associated with ensemble methods like ExtraTrees.

This “Train Wide, Filter Smart” paradigm separates the model’s job — pattern recognition — from the portfolio manager’s job—risk and quality control. It leverag the full power of ML while ensuring the final output is institutional-grade, defensib and robust.

Buy Rules: The Devil’s in the Detail

Building on the “Train Wide, Filter Smart” framework, a crucial technical insight emerges: Not all buy filters are created equal. The effectiveness of a filter is deeply dependent on the normalization method of your underlying AI factors.

Through rigorous backtesting, a clear pattern has emerged:
For “Rank & Date” Normalized Models: Complex percentile filters like FRank(EPS_Revision) > 80 work well. The rank-based system is inherently robust to outliers, making it a suitable partner for other relative, cross-sectional ranking rule

For “Z-Score & Date” Normalized Models: Simple, absolute quality filters like CurFYEPSMean > 0 are dramatically superior. The Z-Score method is highly sensit to distribution tails. Using a FRank filter here often injects the very spiky, extreme values that Z-Scores amplify, leading to volatile performance.

The Underlying Principle:
Your portfolio construction layer shouldn’t fight your feature engineering layer. A FRank filter on a Z-Score model often tries to clean a noisy signal with another noi relative signal. In contrast, a simple quality gate (EPS > 0) creates the stable, well-behaved distribution that Z-Score normalization requires to shine

This explains a key nuance: why a Zscore & Date system with a FRank buy rule can work well on the S&P 500 but fails on small caps? The S&P 500 universe is inherent quality filter; the data is much less noisy to begin with, so the complex rule doesn't fight the normalization.

Why are AI Factor Models robust?

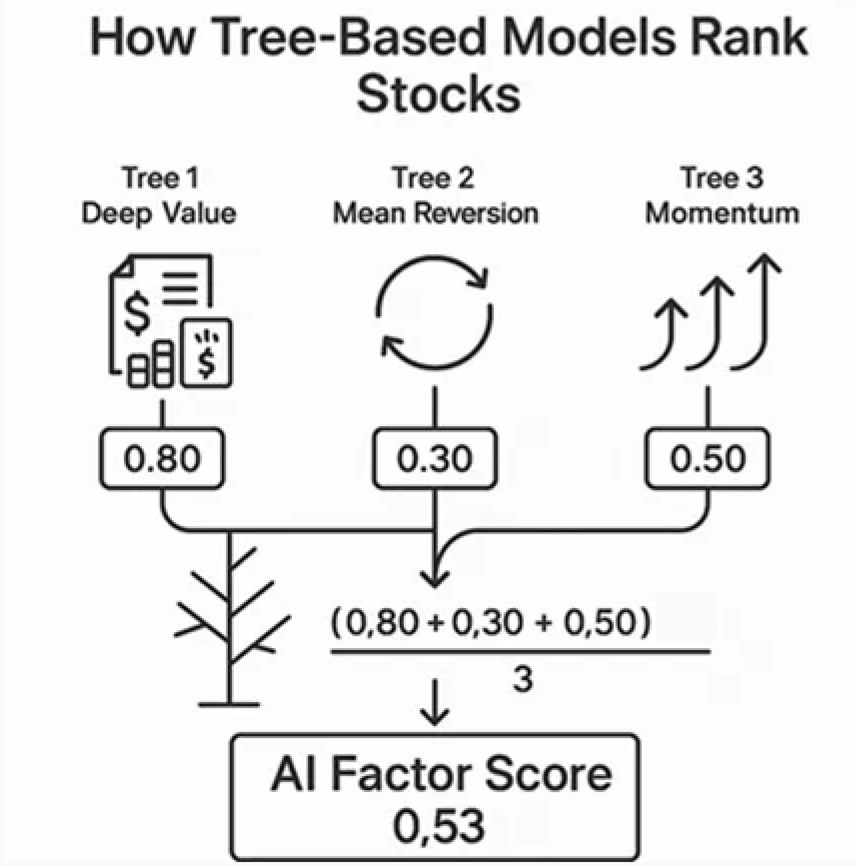
Building a robust AI-driven strategy hinges on a deep understanding of its internal architecture. The true strength isn't in a single "magic" formula, but in a decentralized, multi-layered system designed to withstand market shifts.

Let's break down the engine: an ensemble of decision trees. With a typical setup of features, a depth of 5, and 500 trees, we are not building one model. We are building vast network of micro-rules.

A single tree of depth 5 can have up to 32 leaf nodes — each the end of a unique decision path. The total system capacity is profound:

500 trees x 32 paths/tree = 16,000 potential decision paths.

Factoring in the combinations across 179 features, the number of effective, unique interacting conditions easily surpasses 25,000. These are not complex, hand-crafted rules, but simple, stochastic micro-logic statements.



The core robustness comes from the ensemble method (s above!). Each of the 500 trees is an independent expert, casting a single "vote." The final prediction is a averaged consensus of this entire committee. For a catastrophic failure, a regime shift must invalidate a majority of these 25,000+ paths across hundreds of independent trees simultaneously — a statistical improbability.

AI Factor stays robust, even with buy rules on the portfolio strategy level!

This leads to a critical question: If the model is so complex, why doesn't a strict b rule like "CurFYEPSMean > 0" destroy its subtle intelligence?

The answer is that the buy rule filters the portfolio, not the model's knowledge. The model's 25,000+ rules represent a pre-trained understanding of the entire market landscape. Applying a quality gate doesn't delete this knowledge; it focuses its application on a cleaner, more stable segment.

From the 25,000+ available rules, the model activates a large subset — numbering in the thousands — to rank every stock that passes the filter. The diversity and numbe rules that remain in play are more than sufficient to generate nuanced, intelligent predictions.

We are not reducing a sophisticated brain to a few simple ideas. We are giving it a cleaner dataset to process. The bu rule ensures input stability; the ensemble of trees ensures the output alpha is robust and intelligent.

I have build > 200 Portfolio Strategies with extensive Buy rules, none of them failed OOS Live so far (OOS spanning from weeks to a year).

Example 1

Portfolio Strategy based on an AI Factor System with a Universe of AvgDailyTot(20) < 200000000 (so leaning to the mid and small caps).

Not bad:

LIVE STRATEGIES > UNCLASSIFIED

200 Mil LGBM GPT I

Private Currency: USD

Summary Rebalance Holdings Transactions Trading System Statistics Charts

On 02/28/2022, the model's market exposure deviated 5.04% above the target, an excess of 0.04%.

General Info

PDF Report

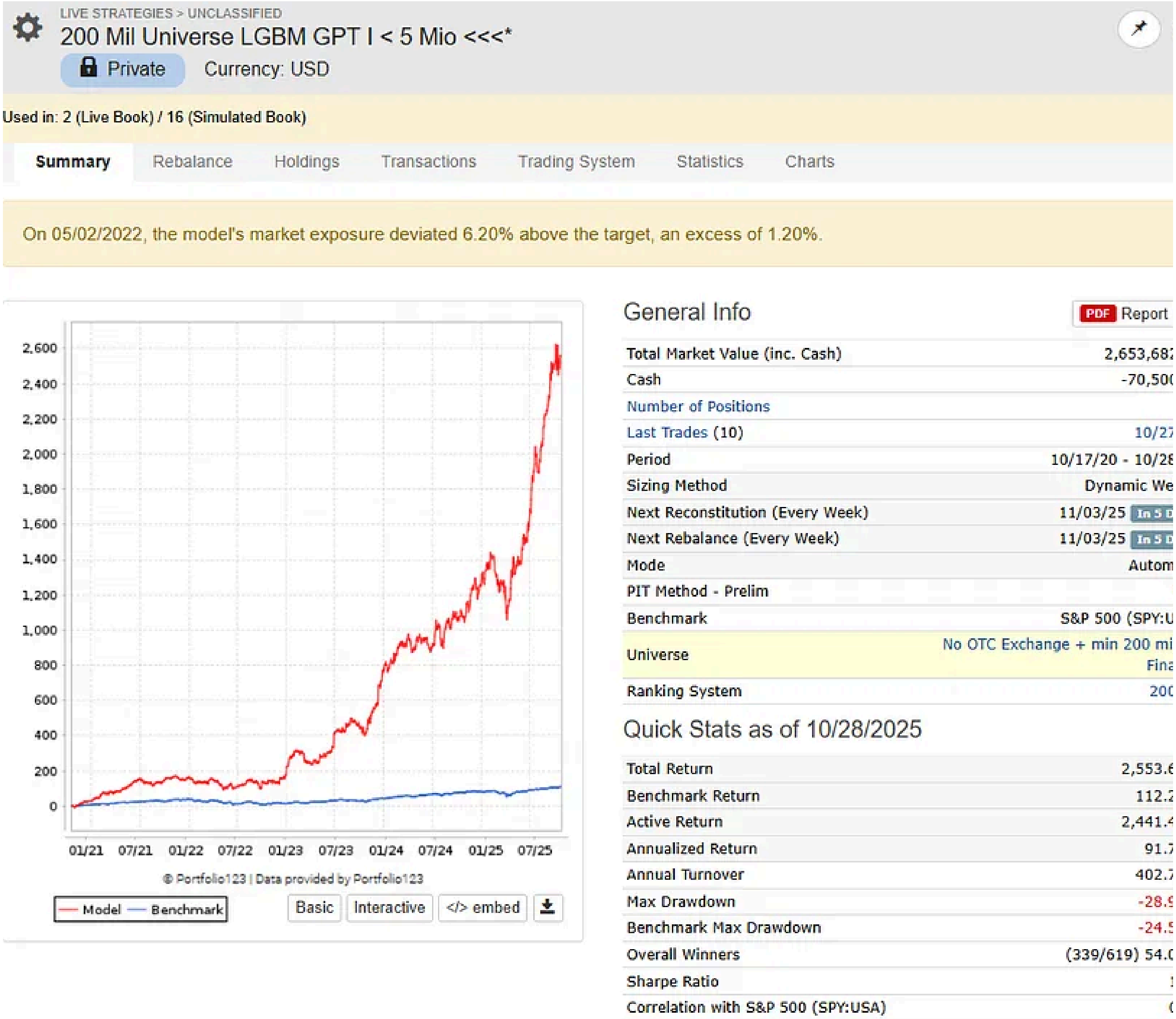
Total Market Value (inc. Cash)	931,61
Cash	-12,61
Number of Positions	
Last Trades (5)	10/2
Period	10/17/20 - 10/2
Sizing Method	Dynamic We
Next Reconstitution (Every Week)	11/03/25 In 5
Next Rebalance (Every Week)	11/03/25 In 5
Mode	Auton
PIT Method - Prelim	
Benchmark	S&P 500 (SPY:U
Universe	No OTC Exchange + min 200 m Fin
Ranking System	20

Quick Stats as of 10/28/2025

Total Return	831.
Benchmark Return	112.
Active Return	719.
Annualized Return	55.
Annual Turnover	278.
Max Drawdown	-29.
Benchmark Max Drawdown	-24.
Overall Winners	(398/712) 55.
Sharpe Ratio	
Correlation with S&P 500 (SPY:USA)	

Now the same strategy with a liquidity filter below 5 Million

AvgDailyTot(20) < 5000000



Example 2

Nothing against it, nice total return strategy:

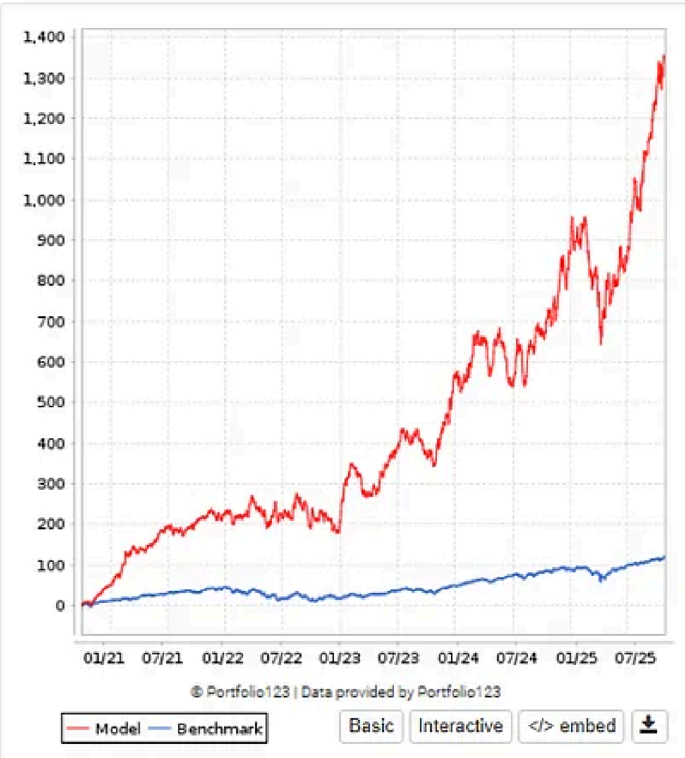
LIVE STRATEGIES > UNCLASSIFIED

3MRELReturn 10 Mil - Zscore Date 178 - NU Lightgbm II - GPT 6/6

Private Currency: USD

Used in: Tree 3 (Simulated Book)

Summary | Rebalance | Holdings | Transactions | Trading System | Statistics | Charts



General Info

PDF Report

Total Market Value (inc. Cash)	1,442,11
Cash	20,11
Number of Positions	
Last Trades (7)	10/2
Period	09/29/20 - 10/2
Sizing Method	Dynamic W
Next Reconstitution (Every Week)	11/03/25 In 5
Next Rebalance (Every Week)	11/03/25 In 5
Mode	Auton
PIT Method - Prelim	
Benchmark	S&P 500 (SPY:l
Universe	No OTC Exchange + min 10 m Fin
Ranking System	3MRELReturn 20 Mil - Zscore Date - NewUniv

Quick Stats as of 10/28/2025

Total Return	1,342.
Benchmark Return	121.
Active Return	1,220.
Annualized Return	69.
Annual Turnover	343.
Max Drawdown	-29.
Benchmark Max Drawdown	-24.
Overall Winners	(468/783) 59.
Sharpe Ratio	
Correlation with S&P 500 (SPY:USA)	

And here comes the kicker, same strategy just with the following rules follows —>

Buy Rules *(Implicit AND)*

copy to screen

Buy1	MedianDailyTot(20) > 50000
Buy3	Rank > 97
Buy3	CurFYEPSMean > 0 or AltmanZOrig > 0

Sell Rules *(Implicit OR)*

copy to screen

Rank	Rank < 80

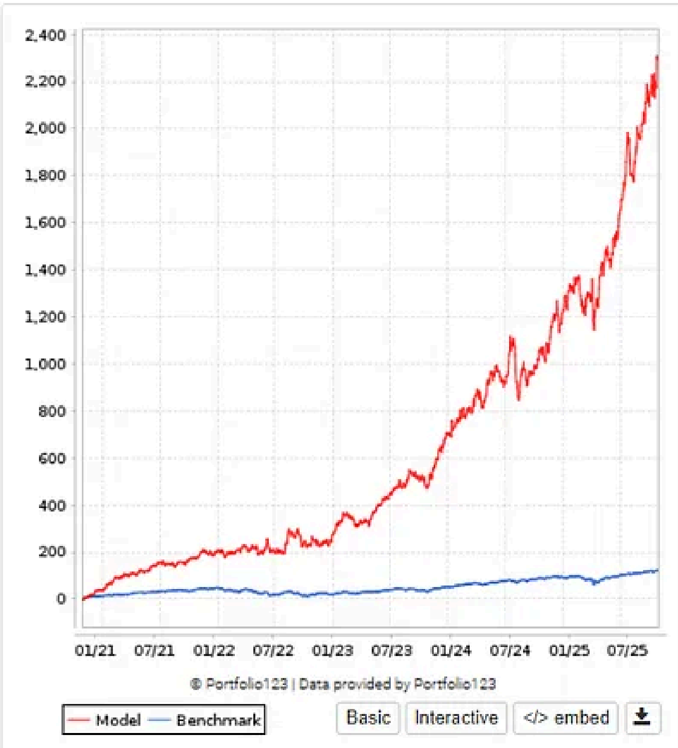
SIMULATED STRATEGIES > UNCLASSIFIED

3MRELReturn 10 Mil - Zscore Lightgbm II - GPT 6/6 ### CurFYEPSMean > 0

Private Currency: USD

Summary | Re-Run | Holdings | Transactions | Trading System | Statistics | Charts

On 07/05/2021, the model's market exposure deviated 5.82% below the target, an excess of 0.82%.



General Info

PDF Report

Total Market Value (inc. Cash)	2,411,049
Cash	123,850
Number of Positions	
Last Trades (2)	10/27
Period	10/28/20 - 10/28
Sizing Method	Dynamic Wei
Last Reconstituted (Every Week)	10/27
Last Rebalanced (Every Week)	10/27
PIT Method - Prelim	
Benchmark	S&P 500 (SPY:U
Universe	No OTC Exchange + min 10 mil Fina
Ranking System	3MRELReturn 20 Mil - Zscore Date - NewUnive

Quick Stats as of 10/28/2025

Total Return	2,311.0
Benchmark Return	125.0
Active Return	2,186.0
Annualized Return	88.9
Annual Turnover	293.5
Max Drawdown	-22.3
Benchmark Max Drawdown	-24.5
Overall Winners	(178/272) 65.0
Sharpe Ratio	2
Correlation with S&P 500 (SPY:USA)	0

While I wouldn't run this as a standalone portfolio, it makes for a powerful satellite allocation designed to boost returns in a larger, core strategy!

Some examples are here: https://x.com/GfI_Himmelreich/status/1981353845260231

The Junk Problem? Yes and No!

It's a common surprise: a sophisticated ML model, trained on hundreds of factors, often load up on speculative stocks. We expect it to learn wisdom, but it only learns patterns.

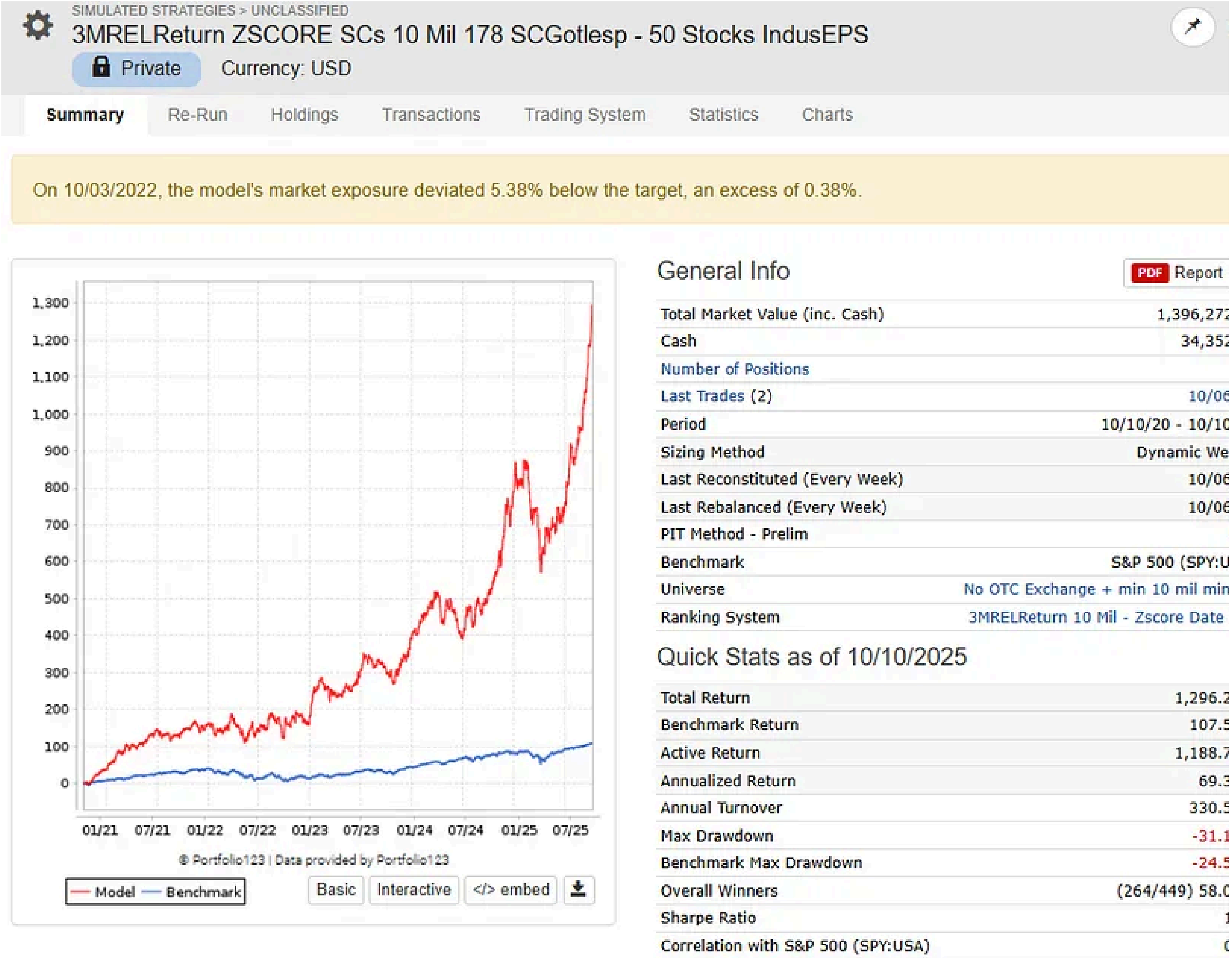
The reason is fundamental. ML optimizes for statistical prediction. Its sole objective is to minimize error, with no innate concept of risk or drawdowns (if the target is price based!).

This leads to the “Junk Factor” problem. And here’s the crucial part: The model is often statistically correct. Junk stocks — with their extreme volatility and binary outcomes — can be highly predictable and immensely profitable: but only during (later) bull markets.

The ML brilliantly identifies these explosive, short-term patterns. It doesn’t understand that they are regime-dependent and come with tail risk.

This is why a pure ML portfolio can be spiky and volatile. It’s chasing patterns that work until they don’t — often spectacularly.

Stuff like this (in 2008 my best guess is, it would have a 70% DD and be up to new highs in 12 Months):



But: I am not routing against those systems, they can be great in a system book!

Why?

Strategies like this, despite their high volatility, can provide valuable uncorrelated alpha and can be extremely effective as a long leg in a market-neutral book or as a satellite allocation to boost overall returns.

Uncorrelated Alpha Source: Its 0.61 correlation to the SPY, while positive, suggests it's not just a leveraged beta bet. It's capturing a different type of risk premia (likely high-octane, small-cap momentum), which is a diversifying return stream.

1. **Ideal for a Long/Short Book:** This is the key. You could pair this high-volatility long portfolio with a different, more stable short portfolio (e.g., shorting low-quality value traps or stable low-momentum stocks). The combined book would aim to be market-neutral, harvesting the pure "junk momentum" alpha while hedging out the brutal market-directional drawdowns.
2. **Satellite Allocation:** In a core-satellite framework, this strategy could be a sma high-conviction "satellite" intended to boost the total return of a much larger, stable "core" portfolio. The core ensures survival; the satellite provides the kick

Some of the best Traders trade Junk ;-)

By the way, it's no coincidence that the best discretionary traders I know — the one who consistently print > 100% years trading breakouts — are masters at knowing exactly when to trade this so-called "junk."

They don't avoid it; they exploit it with impeccable timing and market feel. Their success is the ultimate proof of concept, showing what is possible in these volatile corners of the market.

Train Wide, Filter Smart and weed out th junk (if you want!)

The ML's purpose is raw, unbiased pattern recognition across the entire market landscape — a task it performs with immense scale.

Our role is to apply the economic judgment and risk management it inherently lack By training the model wide and then applying a decisive quality filter at the portfol level, we forge a powerful synergy.

Alternatively, we can consciously choose to harness those "junky" but historically profitable signals, fully aware of their regime-dependent nature. **The key is that it remains our deliberate, strategic decision, not the model's blind directive.**

Best Regards

Andreas

← Previous

Next

Discussion about this post

Comments Restacks

Write a comment...

Brent Harless 19 Dec

Liked by Andreas Himmelreich

So good.

LIKE (1) REPLY