Research Spotlight | Finance, Analytics & Responsible AI

Crash-testing AI portfolio strategies: when Machine Learning shines and when it breaks

Artificial intelligence is often presented as a shortcut to smarter investing: feed a model enough market data, let it learn patterns, and the "optimal" portfolio should follow. But anyone who has tried to translate a promising backtest into a real investment process knows the uncomfortable truth: many strategies look impressive until you change the sample, alter the training window, or move into a new market regime.

In our recent *Applied Soft Computing* publication, researchers affiliated with Luxembourg School of Business set out to test a deceptively simple question that matters to students, teachers, and practitioners alike: **Do Machine Learning–enhanced allocation strategies reliably outperform traditional portfolio methods—or do they succeed mainly under specific conditions?**

Beyond "one perfect backtest"

The paper, *Financial asset allocation strategies using statistical and Machine Learning Models: Evidence from comprehensive scenario testing*, makes a methodological choice: instead of searching for a single winning chart, it stress-tests strategies across multiple scenarios. That matters, because in finance, the boundary between "signal" and "noise" is thin, and small modeling choices can produce big performance differences. To make the exercise realistic, the study uses daily data from 2000 to 2019 for 111 stocks drawn from the NASDAQ-100 and NASDAQ Financial-100, comparing results to a classic benchmark that is difficult to beat consistently, an equally weighted (1/N) portfolio. The research compares six allocation strategies. Four versions of the Mean–Variance framework (including minimum-variance and maximum-Sharpe portfolios in "naïve" and ML-augmented forms) and two Risk Parity approaches, including Hierarchical Risk Parity (HRP), which uses clustering to better reflect how assets move together.

A hybrid approach designed for both performance and explainability

One of the core contributions is a hybrid strategy that integrates forecasting into a familiar portfolio-construction engine. Expected returns are forecast using Facebook/Meta's Prophet, while volatility is forecast using GARCH, an established econometric model designed to capture volatility clustering. Those forecasts then feed into a Maximum Sharpe Ratio allocation (aiming to maximize return per unit of risk). This design is not only about accuracy, it is also about interpretability and compliance, highly complex "black-box" methods may face real barriers in regulated settings, and it highlights the growing importance of transparency.

The headline result, and the caution that comes with it

In the Base Scenario (111 stocks, 2000–2019), the ML-enhanced Maximum Sharpe strategy delivers a result that grabs attention even in a crowded AI-finance landscape. Starting from 10,000 (in monetary units), the portfolio ends at 378,467, which corresponds to 3685% ROI and 19.9% compound annual returns.

In that same scenario, the equally weighted benchmark ends at 229,512 (2195% ROI, 17.0% annual returns). Put differently, the ML strategy's ROI is 1490 percentage points higher than the benchmark in the base case, an outcome the authors emphasize as a best-case illustration of what ML can achieve when conditions align.

But the paper's central message is not "ML wins." It is: ML can win big, but you only earn the right to say that after you test robustness.

When conditions change, performance changes—sometimes dramatically

To probe generalizability, the study runs additional scenarios that modify the training period, the stock universe, and the tuning approach. In one variation, the dataset is extended back to 1995, in another, stocks are included with different missing-data handling and training windows. The results show why scenario testing belongs in every analytics curriculum. In Variation 1, the ML Maximum Sharpe strategy drops to 1448% ROI and 14.7% annual returns (still positive, but far from the base-case standout performance). In Variation 2, it improves to 2308% ROI and 17.2% annual returns, yet still does not replicate the base-case dominance. A particularly instructive point emerges here for students: in these alternative scenarios, the naïve Maximum Sharpe strategy (using simple historical estimates rather than ML return forecasts) becomes exceptionally competitive. It delivers 2411% ROI in Variation 1 and 2913% ROI in Variation 2—outperforming the ML Maximum Sharpe strategy in both cases. It shows that in noisy settings like returns forecasting, "more sophisticated" does not automatically mean "more reliable," and that well-understood baselines deserve respect, especially when regimes shift.

The risk question: higher returns, but not consistently better risk-adjusted outcomes

Another key lesson of the study is that absolute performance and risk-adjusted performance do not always travel together. The ML Maximum Sharpe strategy, while capable of delivering the highest returns in the best-case scenario, is also consistently described as the most volatile strategy across scenarios, with risk-adjusted performance that is not uniformly superior. In contrast, HRP stands out for consistency. The study finds that HRP outperforms naïve Risk Parity in every scenario and delivers the most stable risk-adjusted results over time, often staying above the benchmark in Sharpe ratio terms even when other strategies fluctuate around it. As we might know, HRP also offers a valuable bridge between classic diversification principles and modern ML thinking. It uses clustering to respect correlation structure and avoid hidden concentration in assets that "look diversified" only when viewed one-by-one.

A final note

This research highlights why the most important capability in modern finance is not merely knowing how to fit a model, but knowing how to evaluate it honestly. This could be done through scenario testing, sensitivity checks, and understanding of how training data shapes outcomes.

The most valuable takeaway may be the simplest: the goal is not to build models that look good in one backtest, but strategies that remain credible when reality changes. That is the kind of thinking that turns analytics into decision-making, and research into impact.

Reference

B. Penayo, V. Pribicevic, & A. Novak. (2025). *Financial asset allocation strategies using statistical and Machine Learning Models: Evidence from comprehensive scenario testing. Applied Soft Computing*, 177, 113193.