Large Scale Machine Driven Statistical Arbitrage

Value Proposition

I contribute a strategy that has a **low start-up cost (**monetizing technical and fundamental datasets), **low execution risk**, and a **fast go-to-market time of less than 3 months** (intimately familiar with the infrastructure to run the strategy). **The strategy in its nascent form has generated ~3% returns on a ~100M GMV book at 2 sharpe for 100 production models, and the strategy will scale at log(num models) in sharpe (returns).** The maximum capacity of the strategy is **magnitudes larger than 100M** and depends on the universe of choice. For each universe, **capacity scales at sqrt(num models)** up till the capacity limit of the universe. The number of models scales exponentially with the number of datasets, hence, once existing datasets are exhausted, I can **identify 100+ datasets with high ROI to improve the strategy and add to the team's data pool.**

Introduction

My hypothesis is that we can treat the markets as an abstract problem, and utilize machine-learning principles and theory-free modeling to predict future relative value. Certain datasets provide a glimpse of the future of relative stock returns, and we utilize a suite of algorithms to refine and harness information from these datasets. Our approach thrives on scale, which enhances performance and builds a protective moat against competitors. Furthermore, a fully automated pipeline ensures costly human capital can focus on innovation and refinement.

Search Algorithms

Search algorithms that seek to derive good models are employed on the datasets, and can only utilize some available historical data to find good models. Each model is a weak predictor relying on a few data fields to forecast relative future returns. Good models meet criteria such as high sharpe ratios, low exposure to risk factors, low correlation, robustness to universe sizes, low turnovers, and high capacities.

Theory Free Searches

When running search algorithms, the lack of an imposed structure will trade off type II errors for type I errors. It is inevitable that there will be noise introduced to the production pool, and this is true whether or not the models are found by a machine or a human. However, the core belief here is that our process can find true positive models that produce useful predictions by utilizing datasets that contain information about future returns.

Validation Process

Given the risk of overfitting, the models undergo stringent checks to increase the probability of robustness. A designated hard validation period is kept separate from the search algorithms. Before good models are productionized, they have to pass the validation process, which validates that the model is able to generalize into an out-of-sample period and value-adds to the production pool of models.

Overcoming Noise

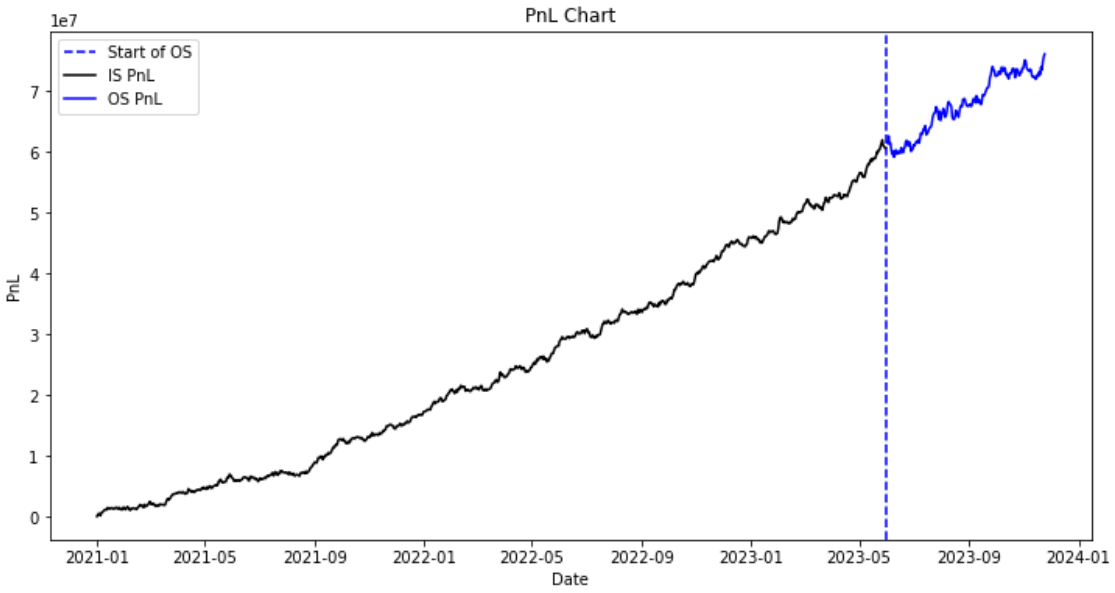
Some models will still be productionized out of sheer statistical chance despite being noisy. The primary insight is that we can also apply the law of large numbers to models. As long as our process can generate models with positive expectations, we can ensemble a large number of models, and the asymptotic behavior of false positive models is that noise signals will cancel each other out. As long as the PnL generated from true positive models can overcome the turnover cost of the noise models, the strategy will be able to generate positive PnL.

Technology

Infrastructure determines the rate at which models can be discovered. We can construct an efficient and scalable infrastructure that backtests ~3000 instruments across 10 years in 10 minutes for a single core. This translates to 150 backtests in a day on a single core, and scales linearly with the number of cores available. **A 64-core, 256GB RAM machine can simulate 10000 models a day.** We can expect a discovery of 10 production models a day at a yield of 1% and hit 1000 production models within 100 days of running a search algorithm.

Summary Statistics

Statistic (OS)	Value
Live Since	2023-06
Region	GLOBAL (USA, ASI, EUR)
GMV	100M
Annual Returns	3%
Annual Variance	1.5%
Annual Sharpe	2
Max Drawdown	0.7%
Turnover	20%



Scaling Capacity & Sharpe

Out-of-sample sharpe scales at the rate of log(N), where N is the number of models, i.e., a 1000 model portfolio will run at a sharpe of ~3 out-of-sample. The 100 productionized models come from 3 datasets, but a comprehensive study has been completed across ~100+ datasets to discover ~2000 candidate models. The production portfolio has a GMV of 100M, and the maximum capacity of the strategy is several magnitudes of that since the universe is ~10000 equities wide. The capacity of the strategy scales at sqrt(num models) up till the limit of the universe after the initial 100M, i.e., a 1000 model portfolio will have a capacity of ~3000M. We target annual returns of 6% and annual volatility of 2% at a sharpe of 3 at scale (1000 models and >500M).

Conclusion

Our approach with machine-driven statistical arbitrage allows us to utilize the law of large numbers at the model level, and extract unintuitive information from datasets about future returns. As the number of datasets and models increases, the number of uncorrelated, true positive signals will increase and so will our performance. Since our process is largely algorithmic, incremental improvements increase marginal benefits and decrease the marginal costs of each new dataset.