CORE MATTERS

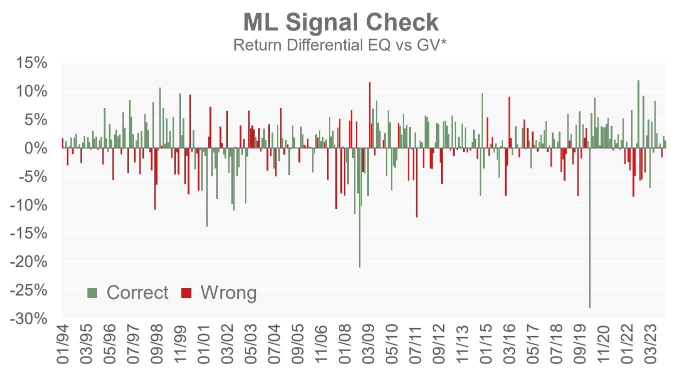
Machine Learning and Tactical Asset Allocation – Part I: Integrating the Machine Learning Signals

Thorsten Runde

February 15, 2024

Our Core Matters series provides thematic research on macro, investment, and insurance topics

- So far, we have been focussing on traditional econometric approaches for modelling economically sound relationships. In contrast, machine learning (ML) employs algorithms to learn patterns directly from data. It can handle large datasets and complex relationships. As such, ML is a promising and flexible supplement to our existing models.
- In a traditional insurance asset management company, the specialists' expertise is the central pillar of the allocation decisions. We here develop a concept to complement the existing human expertise with ML signals to improve the results of our processes.



Source: Datastream, GenAM
*Based on average mid-month data for the United States-Datastream Market Index (EQ) and BBG US Government Bond Index (GV)

- We generate ML signals that forecast the upcoming market regime simply defined as equities outperforming government bonds or vice versa. The signals are generated by an algorithm that systematically searches a large macroeconomic database for comparable situations in the past. These signals amend our existing tactical asset allocation process as an additional input layer. They are used to adjust the overall depth of the initially recommended active positioning, applying simple rule-based overlay strategies.
- In a true out of sample check since 06/2022 a tactical asset allocation (TAA) exclusively based on the machine learning signals would have clearly outperformed its benchmark and even naïve allocation strategies. For the intended use case i.e., complementing our more comprehensive TAA approach, we find in a long-term simulation study that the added value generated by particular combinations of machine learning overlays can be expected to range from 20 to 30 bps per year.

CONTENT

1. Basic idea and motivation

Artificial Intelligence is transforming and redefining the way firms operate and make decisions. It offers a huge potential for the financial industry, including the effective management of portfolios. A particularly interesting area is the use of AI for the tactical adjustment of portfolios to current market conditions.

Very often the two terms artificial intelligence (AI) and machine learning (ML) are used synonymously. Whereas the first in general refers to computer software mimicking human cognition, the latter is just one of various options to implement such a software. Hence, ML can be regarded as a comprehensive state-of-the-art toolbox that can be used to profitably complement approaches from traditional econometrics.

Integrating ML into an already existing and proven TAA process

In traditional insurance asset management, the expertise of the specialists is the central pillar of allocation decisions. Thus, we are convinced that the application of ML can only unfold its full potential when treated as a complement to the human expertise. Therefore, – although also providing value on its own – integrating ML into an already existing and proven TAA process is the focus of our analysis.

In the following we show how ML techniques applied to freely available data can add value to a TAA process of a simplified portfolio consisting of US Treasuries and US equities. We explain how to tailor the ML training setup to ensure a smooth integration into an existing TAA approach. We conclude presenting back-testing results that show the added-value of a ML-enhanced TAA process versus a traditional one.

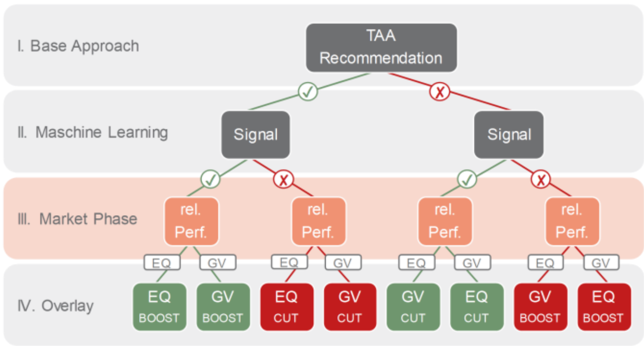
2. Practical application

In general, TAA is aiming at enhancing the returns of a portfolio relative to a benchmark by tactically altering its weighting structure. We do so by applying a Markowitz-like optimisation based on total return forecasts resulting from a top down research approach. We complement the TAA process with the ML signals by adjusting the degree to which the weighting structures of the portfolio and its benchmark deviate.

ML geared towards real-life use case

We aligned the training phase of the ML with the real-life framework conditions and developed strategies on how to combine the ML signals with the recommendations from the existing TAA process.

TAA Overlay Scheme



Source: GenAM

2.1 Implementation

As can be seen in the chart above, we start from our existing TAA process which is labelled as: *I. Base Approach*. From this base approach we get a TAA recommendation[1] which coincides with a concrete portfolio structure i.e., active weights for equities (EQ) and government bonds (GV). In the next layer *(II. Machine Learning)* the ML signals are generated. Like the base approach recommendation, they are also either

[1] For the sake of clarity any kind of TAA recommendation in this analysis just refers to equities outperforming government bonds or vice versa. We focus on just two asset classes as this simplifies back testing (see chapter 4.2 Back-testing the implementation) by eliminating any degrees of freedom in the portfolio construction. Our existing TAA process does of course cover many other additional asset classes.

in favour of EQ or GV. Yet, instead of doing an additional portfolio construction based on these signals we use them in the last layer *(IV. Overlay)* to tweak the recommended active positions resulting from the base approach.

More specifically, we will either be more aggressive in our allocation stance, i.e., raise active tactical weights, if the recommendation from the base approach and the ML signal point in the same direction; and reduce the aggressiveness if they point in opposite directions[2].

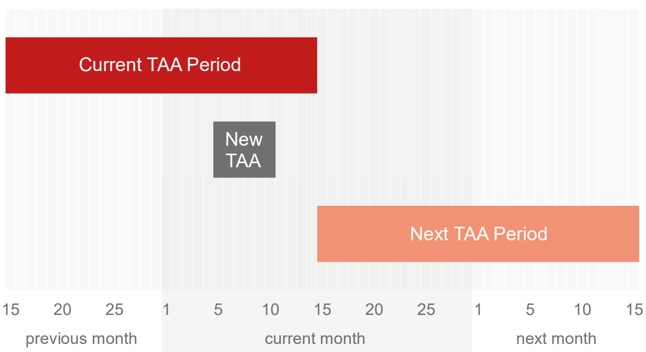
There are four overlays existing. Depending on the realised market phase *(III. Market Phase)*, which is also either in favour of EQ or GV, each of these overlays can either add *(green boxes in the last layer)* or destroy *(red boxes)* value relative to the base approach. "BOOSTing" overlays raise the aggressiveness of the underlying TAA stance while "CUTting" overlays reduce it. A BOOSTing overlay can only add value if applied to a correct TAA recommendation, otherwise it will destroy value. The opposite is true for CUTting overlays[3].

The correctness of both the TAA recommendation and the ML signal are of course unknown when applying any of these overlays. That said, we can make some reasonable assumptions about the probability of its correctness as well as the correctness of the ML signal[4]. Probabilities of occurrence of specific market phases (EQ or GV) can also be derived from historical data[5].

2.2 Organizational framework

We design the ML signals in a way that mimics the conditions under which the TAA recommendations of the base approach are formulated.

Stylised GIAM TAA Cycle



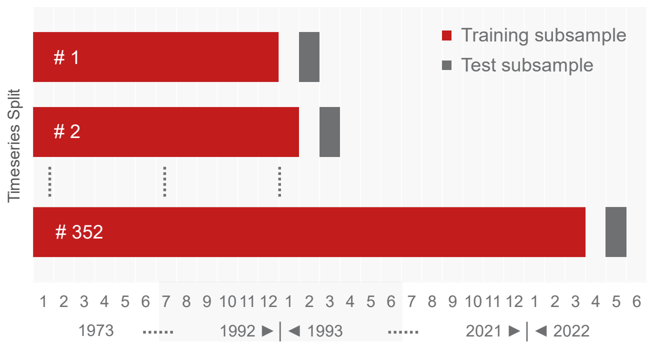
Source: GenAM

In the internal process, the regularly updated TAA recommendations refer to a period of one month length, starting by mid-month (see "Next TAA Period" in the chart on the left). The new recommendation ("New TAA") is made a few days before the implementation period starts. Furthermore, the period of the previous recommendation ("Current TAA Period") is still not finished. This of course has a direct impact on how the data enters the ML model[6] and on how the whole training of the model is set up.

3. Training the model

When training any kind of model, we generally distinguish between training and test subsamples. Based on the former we let the model learn a set of parameters (in our case e.g., the number of neighbours to be considered[7]). The latter is used to evaluate the model's quality by doing out-of-(sub)sample forecasts i.e., using unknown data. The distinction between training and test subsamples is important to ensure that the model is actually learning and not just memorizing.

Crossvalidation Scheme



Source: GenAM

In the training phase of the ML model, we consider the availability of the input data by applying an appropriate lag, as discussed above. Furthermore, we also consider the missing information about the result of the "Current TAA Period" by establishing a gap with the length of one month between the training and test subsamples. I.e., we prevent the model from using the data of the month preceding the forecast as in the real-life application we would not have the information on the market outcome of that month either.

[2] This is achieved by simply multiplying the active positions from the base approach with a factor larger or smaller than one.
[3] Example: EQ BOOST applies if layers I. and II. both point towards EQ. If the true market phase from layer III. is also EQ, layers I. and II. were correct *(green tickmarks)* and thus EQ BOOST adds *value (green box far left)*. In case layer III. is in favour of GV, layers I. and II. were wrong (*red crosses)*. Hence, EQ BOOST destroys value *(red box far right)*.

[4] The former can be derived from our own archived data, the latter corresponds to the model's accuracy (see chapter 3.2 The final specification).
[5] See chapter 4.2 Back-testing the implementation
[6] The TAA recommendation is made at the beginning of month t. The latest available input data is from the end of month t-2. The TAA period ends mid-month t+1. Thus, we are left with an effective lag of 2.5 months.
[7] See chapter 3.2 The final specification

Training setup aligned with organizational framework

We start with an initial training subsample of twenty years beginning in 01/1973. We let the training subsamples grow by each month as we move in time with our forecasts towards the most recent observations. All in, we repeat this process of training and forecasting a good 350 times (see right chart on the previous page). We finally choose the model parameterization that worked best on average across all these steps.

3.1 The input data

The input data we use is the FRED-MD database[8], which is a freely available database containing monthly economic data for the US going back to 1959. The database is provided and maintained by the Federal Reserve Bank of St. Louis (see box on the right). For the market phases / regimes we rely on total return indices from Datastream[9].

Before being used in the model's training phase the raw data goes through a number of pre-processing steps.

Data Preprocessing Steps*

*Steps 6 and 7 performed within the enhanced hyperparameter tuning

Source: GenAM

For the first three steps (see chart above) we rely on MatLab code which is also provided by the Fed of St. Louis. Variable exclusion and smoothing[10] are performed before the training whereas scaling and dimension reduction are part of the training via an "enhanced" form of so-called hyperparameter tuning[11]. In fact, most algorithms assume that features vary on comparable scales. The same is true for principal component analysis (PCA) which is a standard approach to perform the subsequent "Dimension Reduction" step. Therefore, particularly "Scaling" and "Dimension Reduction" are integral components of any data pre-processing sequence in ML.

The FRED-MD database

The FRED-MD is a large database covering month-end macroeconomic data for 127 economic time series ranging back to January 1959. It is provided and maintained by the Federal Reserve Bank of St. Louis. Monthly updates can be downloaded for free. It covers eight economic areas:

1	Output & Income (16 time series)
2	Labour Market (31)
3	Housing (10)
4	Consumption, orders and inventories (10)
5	Money and credit (13)
6	Interest and exchange rates (22)
7	Prices (20)
8	Stock market (5)

It is designed for the empirical analysis of "big data." The timeseries are updated in real-time through the FRED database. They are publicly accessible, facilitating the replication of empirical work. And they relieve the researcher of the task of incorporating data changes and revisions (a task accomplished by the data desk at the Federal Reserve Bank of St. Louis).

3.2 The final specification

We evaluated various algorithms[12]. The best one turned out to be the k-Nearest-Neighbours (kNN) algorithm, a so called supervised, instance-based classifier. The kNN algorithm searches input data from the past for **observations as similar as possible to the current situation** and derives the forecast as a majority vote amongst these observations with respect to the market phase associated[13].

Apart from the core parameter of the algorithm itself i.e., the number of past observations (neighbours, with eleven being

[8] **F**ederal **R**eserve **E**conomic **D**atabase – **M**onthly **D**ata. https://research.stlouisfed.org/econ/mccracken/fred-databases/
[9] We use US-DS-Market index (TOTMKUS) for Equities and Bloomberg U.S. Government index (LHGOVBD) for Government Bonds, both available from 1973 onwards. We use mid-month total returns, with mid-month represented by the average index value ranging from the 14th to the 16th of each month.
[10] A moving average over two months proved most rewarding.
[11] Hyperparameter tuning refers to finding the optimal parameters for the algorithm in the training phase. As scaling and dimension reduction are not part of the algorithm itself, we talk about „enhanced" hyperparameter tuning.
[12] Amongst others: Support Vector Machine Classifier, Random Forest Classifier, ADA Boost Classifier etc.
[13] For more details see our prospective Core Matters: Machine Learning and TAA – Part II: Decomposing the ML Signals

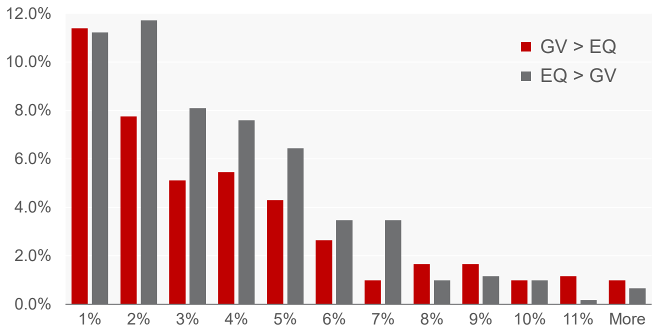
the optimal number) to be considered, further specifications within the pre-processing steps helped to improve the results.

In particular, excluding various financial market areas from the analysis ('Interest and exchange rates' except for the 1-Year Treasury Rate, 'Prices', and 'Stock market' except for the Volatility Index)[14] proved rewarding. Furthermore, the analysis was based on nine principal components derived from the data instead of being based on the original set of eighty-two variables.

Importantly, we not only employed the accuracy as a criterion for the optimal parameterization, but also weighted it with the corresponding return differential. This tweaks the model from just making as many correct forecasts as possible to also **focussing on forecasts in periods where the potential added value from a correct choice is particularly large**.

Monthly Absolute Return Differential

relative frequencies between 03/1973 and 10/2023



Source: Datastream, GenAM

In fact, as the chart above shows, the largest return differentials are more likely to be observed in case GV outperform EQ. Generally, this is true if equity markets crash.

4. The results

Over the period of the model training Equities outperformed Government Bonds in roughly 60% of the observed months. Thus, to achieve an accuracy corresponding to that relative historical frequency, it would be sufficient to naïvely "bet" on Equities throughout. We use this as a kind of benchmark to contrast the results from the kNN approach.

Some Metrics

		Approach	
Metric	Definition	Machine Learning	Naïve
Accuracy	$acc = \frac{TP+TN}{TP+TN+FN+FP}$	63%	60%
$Recall_{EQ}$	$rec_{EQ} = \frac{TP}{TP+FN}$	78%	100%
$Recall_{GV}$	$rec_{GV} = \frac{TN}{TN+FP}$	40%	0%
balanced Accuracy	$bacc = \frac{rec_{EQ}+rec_{GV}}{2}$	59%	50%

Source: GenAM

Hence, the accuracy of the naïve approach is exactly 60% (see Table above). By definition, this approach is always right if Equities actually outperform ($Recall_{EQ}$ = 100%) and at the same time always wrong if Government Bonds outperform ($Recall_{GV}$ = 0%)[15]. Hence, the balanced accuracy must be 50%. By contrast, the kNN approach renders a somewhat lower $Recall_{EQ}$ of 78% but a much higher $Recall_{GV}$ (40%) and thus a higher balanced accuracy of 59%[16].

> *The correct timing of the signals makes the difference*

Thus, the metrics of the kNN algorithm are superior to that of the naïve approach[17], they are less dispersed, and most importantly, the ML model is able to reap the benefits when Government Bonds outperform. This helps to lock in quite some outperformance on the Government Bond side and that is, where the "big points" can be made.

4.1 The signal check

In fact, as can be seen from the upper chart on the next page, the model was correct in forecasting the three periods with the largest return differentials, all of them with Government Bonds outperforming Equities[18]. Eight out the ten largest absolute return differentials occurred on the GV side. Also, seven out of the ten largest points made by the model were made on the GV side. Over the past thirty years an active

[14] Of all things, just excluding the areas ‚Interest and exchange rates' and 'Stock market' might appear counterintuitive, as these areas are considered to be the closest to the market regimes themselves. In fact, excluding them is somehow like rejecting the hypothesis of an adaptive formation of expectations on markets. However, the average regime length is just around two months, with 50% of the regimes only lasting one month. It turned out that taking the past month's regime as a forecast for the next one is less successful than flipping a coin.

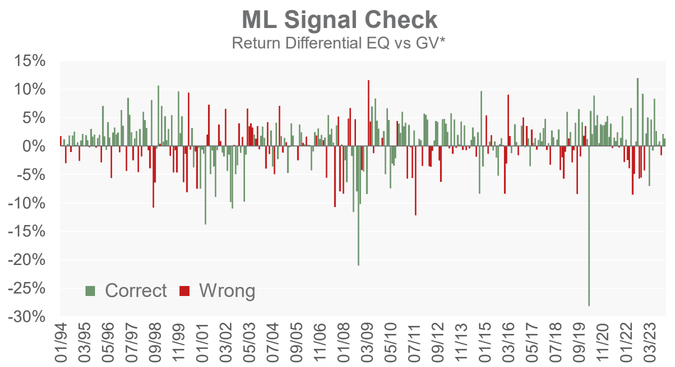
[15] TP = True Positive (Forecast = EQ / Outcome = EQ); TN = True Negative (GV / GV); FP = False Positive (EQ / GV); FN = False Negative (GV / EQ)

[16] ML figures are derived from the monthly model forecast from 02/93 to 05/22, naïve figures from the true market phases.

[17] Of course, this does not hold for the $Recall_{EQ}$. Particularly over such an extended period of time it is unrealistic to generate a rate of 100%.

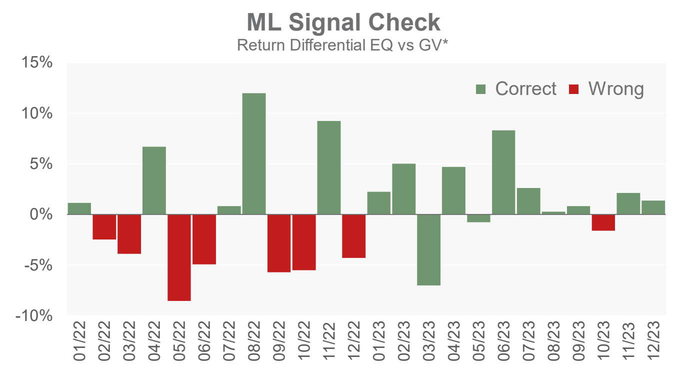
[18] 03/20: -28%; 10/08: -21%; 03/01: -14%

allocation based on the model signals would have been roughly 2.5x as successful[19] as following the naïve approach.



Source: Datastream, GenAM
**Based on average mid-month data for the United States-Datastream Market Index (EQ) and BBG US Government Bond Index (GV)*

Encouragingly, over the new high inflation episode characterizing the past two years the superiority of the model did not change substantially (2.2x). Although the $Recall_{GV}$ shrank to 20% over that period the $Recall_{EQ}$ was at 100%. The model was correct throughout the long market phase in favour of Equity (06/23 – 09/23), a historically rather rare event. In fact, over the past 45 years only 20 out of 134 Equity regimes reached a length or four months or longer[20]. Furthermore, from 06/2022 onwards the signals were generated fully out-of-sample i.e., based on data completely unknown to the model so far. In 2023 the model was correct in eleven out of twelve months (see chart below). Most remarkably, the permanent regime switching from 02/23 to 06/23 was forecasted correctly.



Source: Datastream, GenAM
**Based on average mid-month data for the United States-Datastream Market Index (EQ) and BBG US Government Bond Index (GV)*

[19] Relative performance vs. a 50/50 benchmark; (just for the purpose of this simple comparison).
[20] With nine out of 134 the corresponding ratio for GV regimes is not even half as large.
[21] See chapter 2.1 Implementation

4.2 Back-testing the implementation

The previous statistics show that the stand-alone use of the ML signals renders very promising results. In a next step, we evaluated the results from their integration into the existing TAA process. We assess the added value to a base TAA process by complementing it either with a single overlay or a combination of overlays[21].

The challenge here is that for our internal 'base' TAA process, we do not have adequate recommendations reaching back thirty years into the past. Given its heuristic character it is also excluded to reproduce the missing recommendations ex-post. But, from the track record of our TAA process, which goes back to 2003, we can derive an accuracy rate comparable in content to that of the ML model. With 60% this rate turned out to be also similar in size.

Added Value vs Base Approach
in bps p.a.

	TAA Enhancements	Simulation Quantile 0.5%	2.5%	5.0%	median	95.0%	97.5%	99.5%
Machine Learning	EQ BOOST	12	13	14	17	21	21	23
	GV BOOST	2	4	4	8	12	12	13
	EQ CUT	-5	-3	-3	1	5	5	7
	GV CUT	-1	0	0	4	7	8	9
	EQ & GV BOOST	19	21	21	25	29	30	32
	EQ & GV BOOST + GV CUT	23	24	25	29	33	33	34
Naïve	EQ BOOST	9	11	12	15	19	21	22
	EQ OW (vs BM)	*-74*	*-63*	*-57*	*-26*	*7*	*14*	*27*

Source: GenAM

We used this historical rate to simulate monthly 'base' TAA recommendations over thirty years. We did so via 10.000[22] simulations ensuring that the accuracy in each of the simulation runs exactly meets the derived accuracy of 60%. Just the dates differ where the simulated recommendations are correct/wrong. We use the ML signals and the simulated TAA recommendations to construct model portfolios for a simple 40/60 US equity/bond benchmark[23]. In doing so, we do not only get an idea of how the ML-enhanced approaches perform over time compared to the base approach but also of how the relative performance is distributed around it. The table above summarizes the results for all single and the most successful combined overlay strategies.

As can be seen the most successful strategies are a combination of the two overlays EQ & GV BOOST (which trigger

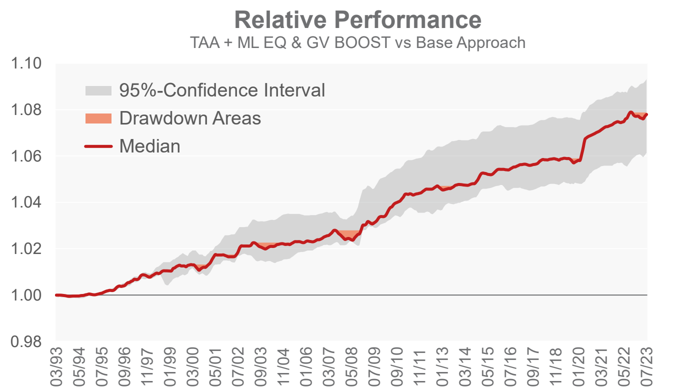
[22] In fact, we did so 10.001 times to guarantee a unique assignment between the quantiles of the distribution of the relative values and the simulation runs.
[23] TAA setup: Effective test period (03/93 – 07/23); Benchmark: 40% US EQ / 60% US GV; allocation range: ±10 pp; monthly rebalancing; no neutral positioning; full exhaustion of allocation range; Overlay integration: 1.25x allocation range (BOOST); 0.75x (CUT)

stronger active positions when the ML signal supports the base recommendation) and the three overlays EQ & GV BOOST & GV CUT (which additionally cuts active positions if the ML signal diverges from the base recommendation in favour of GV). These overlay strategies generate added values between ~20 and ~35 bps with a confidence of 99%[24].

> *A combination of the two BOOSTing overlays is the strategy of choice*

The success of these strategies is not too surprising. They are the ones with the highest net probabilities of adding value.[25] These probabilities are determined by three factors: (1) the quality of the base approach, (2) the quality of the ML model, and (3) the likelihood of EQ outperforming GV.

The strategy of combining the two BOOSTing and the GV CUT overlay appears to be the superior one in terms of relative performance and also in terms the of net probability of adding value (29.2% vs. 23%). However, our recommendation is to stick to the simple combination of the EQ & GV BOOST overlays.



Source: GenAM

One must keep in mind that the performance quantiles shown in the table on the previous page were derived over an exceptionally long investment period. In the short run the dispersion of the results will be higher. As the chart above shows for our preferred strategy, the trend outperformance is interrupted by drawdown episodes, i.e., periods in which the ML enhanced TAA can fall back behind the Base Approach, of more than one year in length and -46 bps in depth.

For the strategy also including the GV CUT overlay, although being the superior one over the complete period of thirty years, these drawdown periods may even be twice as long and one and a half times as deep. The occurrence of drawdown periods lasting longer than one year is also more than twice as high.

Furthermore, for CUTting overlays the accuracy of the underlying ML signals must be larger than that of the base approach to reach a positive net probability of adding value. CUTting can intuitively be interpreted as overruling. It is quite natural to require the one approach overruling the other to be the more accurate one. For BOOSTing strategies, this hurdle is distinctively lower. It can be shown that one minus the accuracy of the base approach already represents a sufficient level of quality of the underlying ML signals.

5. Conclusions

This study shows that it is possible to successfully improve tactical portfolio decisions with ML signals boosting the performance of existing decision processes.

We showed that for an existing TAA process that renders 60% success rate in tactical bond/equity allocation, the outperformance can be boosted by increasing active positions when the ML signals support the TAA recommendations derived from the independent base TAA approach. In a simulation study such a combined approach was able to add 20 to 30 bps p.a. to the base approach.

In the training phase of the model there is no mathematically well-defined and behaving objective function that can be maximized like in traditional econometric approaches. Instead, the learning process of the machine is based on exploring all possible combinations of model parameters, selected variables, and pre-processing measures. These can easily add up to hundreds of thousands. Testing all of them quickly runs into computational limitations. While there may be even superior possibilities, our model is sufficiently accurate, maintainable at reasonable costs, and open to further analyses of the ML signals themselves[26].

In this seminal study, we focused on the US and the return differential between Equities and Government Bonds. There is broad leeway for extending the scope of the investigation e.g., via (a) a shift of the regional focus towards Europe and (b) an inclusion of further asset classes like Corporate Bonds.

[24] The figures are based on a time period of thirty years. They might differ substantially for shorter investment periods.

[25] E.g., EQ BOOST: Let us assume independence between the layers of the TAA Overlay Scheme (see 2.1 Implementation). BOOSTing in general only adds value if TAA recommendation and ML signal are correct. The probability for this to happen is 60% x 63% = 37.8%. Additionally, for EQ BOOST to add value Equities must outperform Govies: 60% x 37.8% = 22.7%. If Govies outperform (1 - 60 % = 40%), the TAA recommendation (1 - 60 % = 40%) as well as ML Signal (1 - 63% = 37%) falsely point towards Equities and EQ BOOST will destroy value: 40% x 40% x 37% = 5.9%. Thus, the net probability of adding value is 22.7% - 5.9% = 16.8% (GV CUT: 6.2% and EQ CUT: -3.2%).

[26] See our prospective Core Matters: Machine Learning and TAA – Part II: Decomposing the ML Signals

Imprint

Issued by:	**Generali Asset Management S.p.A.** **Società di gestione del risparmio, Research Department**
Head of Research:	**Vincent Chaigneau**
Head of Macro & Market Research:	**Dr. Thomas Hempell, CFA**
Team:	**Elisabeth Assmuth |** Research Operations **Elisa Belgacem |** Head of Cross-Asset Quant & Dev, Senior Credit Strategist **Radomír Jáč |** GI CEE Chief Economist **Jakub Krátký |** GI CEE Financial Analyst **Michele Morganti |** Head of Insurance & AM Research, Senior Equity Strategist **Vladimir Oleinikov, CFA |** Senior Quantitative Analyst **Dr. Thorsten Runde |** Senior Quantitative Analyst **Dr. Christoph Siepmann |** Senior Economist **Dr. Florian Späte, CIIA |** Senior Bond Strategist **Guillaume Tresca |** Senior Emerging Market Strategist **Dr. Martin Wolburg, CIIA |** Senior Economist **Paolo Zanghieri, PhD |** Senior Economist
Head of Insurance and AM Research:	**Michele Morganti**
Team:	**Raffaella Bagata |** Research Operations **Alberto Cybo-Ottone, PhD |** Senior Economist **Mattia Mammarella |** Research Analyst **Roberto Menegato |** Senior Insurance Research Analyst **Antonio Salera, PhD** | Economist, Pension Expert **Federica Tartara, CFA |** Senior Economist
Head of Credit Research:	**Vivek Tawadey**