ROBECO
The Investment Engineers

WHITE PAPER

From black box to glass box:

Understanding and attributing machine learning models

Contents

Introduction 3

Machine learning at Robeco 4

Understanding machine learning predictions 5

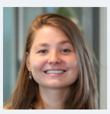
Attributing machine learning portfolio returns 9

Conclusion 11

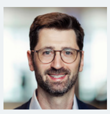
WHITE PAPER JUNE 2025

Marketing material for professional investors, not for onward distribution.

Tobias Hoogteijling
Quant Researcher



Vera Roersma
Quant Researcher



Matthias Hanauer
Executive Director
Quant Research

Introduction

Machine learning (ML) techniques are increasingly important in quantitative investing thanks to their ability to capture complex financial data dynamics and improve return and risk estimates.[1] Algorithms such as neural networks and gradient-boosted regression trees can capture nonlinear relationships and interactions in a flexible way. The resulting higher complexity means they are sometimes perceived as being 'black boxes'. However, many tools exist to understand and interpret machine learning models. In this white paper we aim to provide insights into such tools, focusing on both predictions and strategy return attribution.

As we embrace these more advanced modeling techniques as part of our next-gen quant efforts,[2] it is essential to foster an equally advanced understanding of the data and models. Publicly available tools, supplemented with proprietary techniques, help us interpret ML predictions and attribute performance, transforming the 'black box' into a 'glass box'.

Machine learning models

- ML models add value in stock selection strategies
- Rationalizing ML algorithms is vital for their successful use and clients' trust
- We apply proprietary techniques to understand the drivers of ML predictions and to attribute the performance o f ML strategies

1. Numerous academic studies show that models allowing for nonlinearities and interactions, such as neural networks and tree-based models, lead to superior out-of-sample (OOS) returns compared to linear models. For instance, Gu et al. (2020) predict stock returns for the US, Tobek and Hronec (2021) for developed markets and Hanauer and Kalsbach (2023) for emerging markets.
2. Learn more about next-gen quant on our dedicated page on the Robeco website: Next-generation quant.

Machine learning at Robeco

Over the past years, a wide variety of ML signals have been added to Robeco's quant strategies. These signals can be related to returns or risk; span short-term to long-term horizons; and function as either direct alpha signals or indirect idea generators. Some models rely solely on numerical inputs, while others also incorporate other types of data, such as textual or audio information.[3]

In this white paper we aim to provide insights into the tools we use to understand and interpret ML models, focusing on both predictions and strategy return attribution.

How to train machine learning models

ML models are trained to estimate the relationship between stock and company characteristics (also referred to as features in ML nomenclature) and subsequent returns. To this end, various classes of algorithms exist. These models differ in how they structure this relationship, comprising both simpler, linear models and more complex models that allow for nonlinearities and interactions.

Examples of linear models are Ordinary Least Squares (OLS) and penalized linear models such as Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (ENET). Models that allow for nonlinearities and interactions include tree-based methods such as Random Forests (RF) and Gradient-Boosted Regression Trees (GBRT), and Neural Networks (NN).

However, various design choices exist for a given model, including input features and their transformation, the exact definition of the dependent variable (the target), the training sample, details of the model retraining, and the choice of algorithm hyperparameters (cf., Chen et al., 2024). The optimal choices depend on the desired use of the ML predictions and require both econometric and quantitative investing (domain) knowledge.

To showcase some of our available tools, we investigate various Gradient-Boosted Regression Tree models.

3. For a broader view of the different applications of ML in asset management, we refer to Blitz et al. (2023), Chen et al. (2023), and Chen and Zhou (2023).

Understanding machine learning predictions

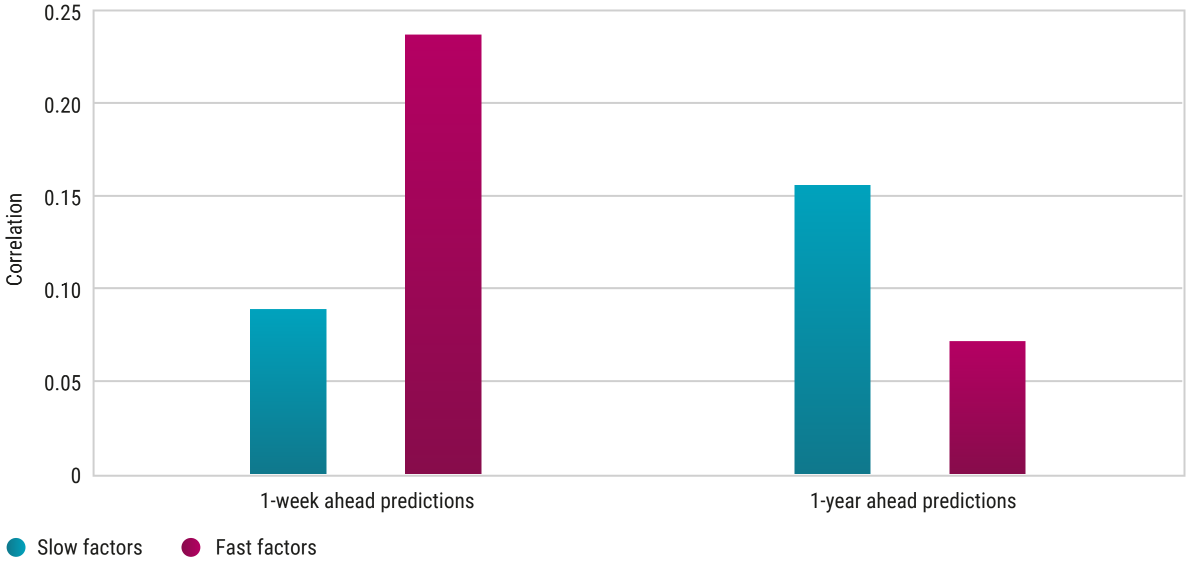
To identify the factors that influence an ML algorithm's predictions, we typically analyze how the predictions vary when the input features are adjusted. The stronger the prediction responds to the variation of an input feature, the more important it typically is. A very simple analysis is examining correlations between ML predictions and the underlying input features. The outcome of this analysis should validate our initial hypotheses, which are grounded in economic rationale.

For example, consider two ML models: one that predicts stock returns one week ahead and one that predicts returns one year ahead. Given that these models are trained to predict returns over distinctly different horizons, we expect them to pick up different patterns in the data. Specifically, we expect the one-week-ahead model to relate more strongly to 'faster' signals such as short-term reversal or short-term momentum, while the one-year-ahead model relates more strongly to 'slower' signals, such as value and quality. The average correlations between the predictions of the two models and a set of fast and slow equity factors are shown in Figure 1.

We observe that the one-week model exhibits a higher correlation with fast factors than slow ones, while the one-year model shows the opposite pattern. This example illustrates the data-driven nature of ML algorithms: the algorithm emphasizes different features depending on the prediction target. Thus, the model training process responds to variation in the prediction target in the way it was expected.[4]

A more sophisticated analysis of ML predictions can be performed based on Shapley values (Shapley, 1953). Shapley values describe what part of a prediction is due to which specific input feature, thereby taking into account not only the linear effect of a feature on predictions, but also nonlinearities and interactions (for more details, see the text box on page 6). One of the key advantages of Shapley values is that they are additive, meaning that they allow us to break down predictions into the individual contributions of each feature used in the model.

Figure 1 | Correlations of machine learning predictions with known equity factors

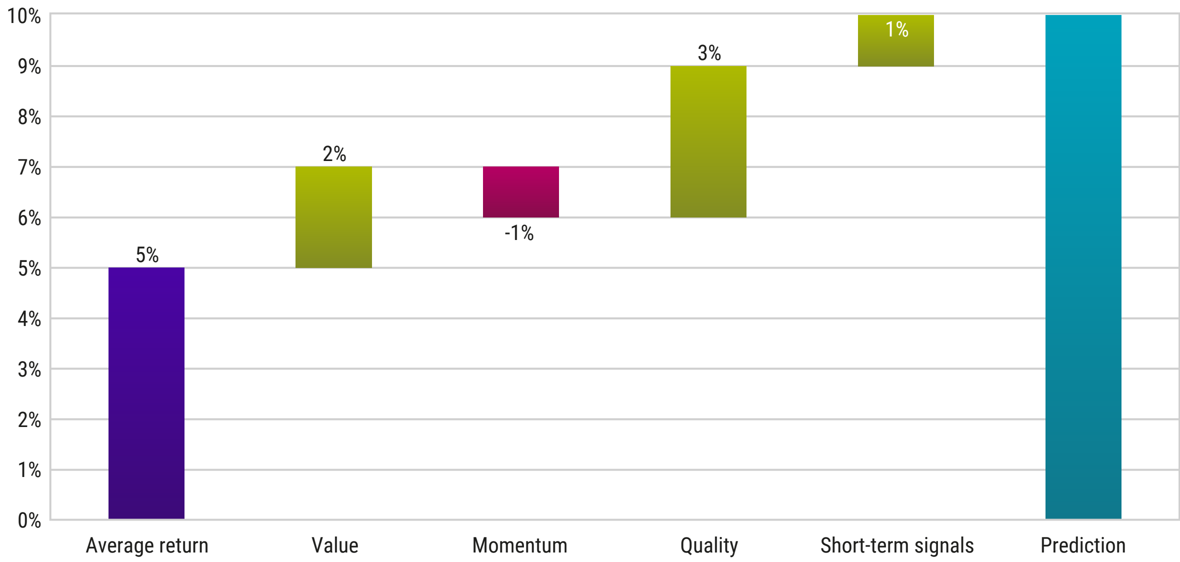


Source: Robeco. The figure shows the average cross-sectional correlation of predicted returns from models trained on stock returns one week and one year ahead with both a slow and fast composite signal score. The model is evaluated from January 2015 to April 2025, using constituents of the MSCI World Index. The results shown are hypothetical, based on theoretical models and assumptions, and are no guarantee of future results.

4. We describe the relationship between forecast horizon, model characteristics and turnover in more detail in the paper entitled "The term structure of machine learning alpha" by Blitz et al. (2023).

An illustrative example is shown in Figure 2. In this case, the average predicted return across all stocks is 5%. For the specific stock considered, the prediction is 10%. Specifically, value, quality, and short-term signal features contribute positively to its higher-than-average prediction, while momentum contributes negatively. This breakdown varies across stocks and is shaped by different interactions and nonlinearities.

Figure 2 | Illustrative example of Shapley prediction decomposition



Source: Robeco. For illustrative purposes only. The figure shows a hypothetical example of how a final machine learning model prediction can be decomposed into underlying feature components.

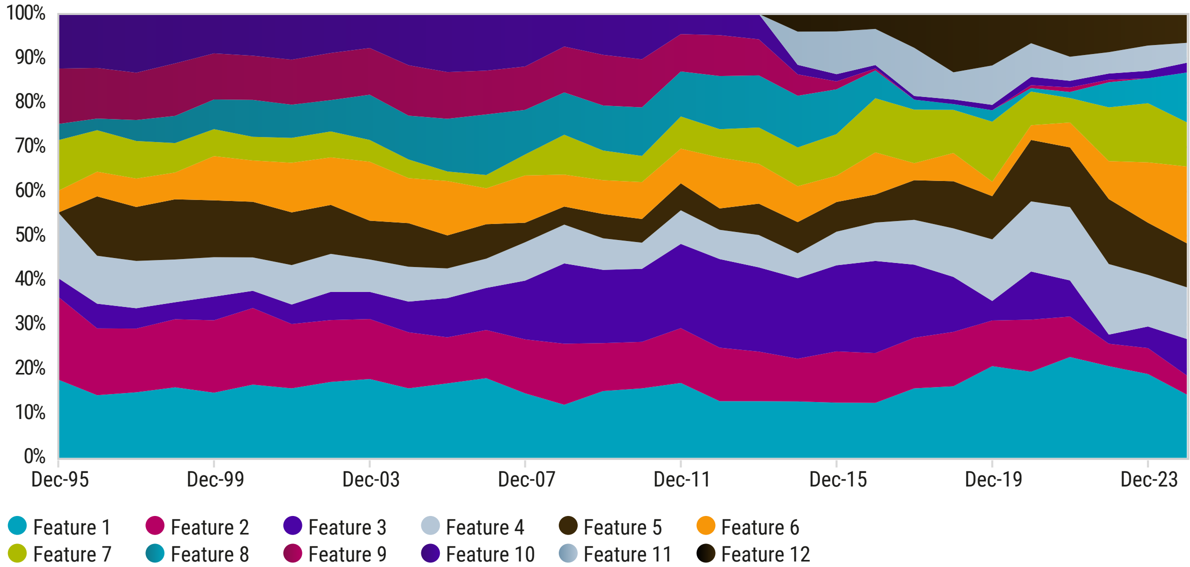
Shapley values versus linear regression coefficients

In a sense, Shapley values are akin to linear regression coefficients, as both describe how much predictions change in response to variation in the explanatory variables. However, regression coefficients are the same for all observations. Shapley values on the other hand, allow for a different effect of features on predictions for each observation. This heterogeneity is the result of interactions and nonlinearities.

Because Shapley values are additive, we can aggregate them across all security predictions to determine which features are the most relevant overall. Figure 3 shows the aggregated absolute Shapley values over time for a one-month-ahead stock return prediction model, displayed for the twelve most important features. The area occupied by each feature indicates its importance for the final model prediction.

The plot illustrates how feature importance evolves over time, with the model adapting gradually to new data. Feature 1 holds the greatest importance and remains important throughout the entire period. In contrast, Feature 2's importance diminishes over time. Features 11 and 12 appear later in the sample, so their importance is only visible from 2013 onward. In fact, it seems that they largely replace Features 9 and 10 from 2013 onward. As with correlations, the Shapley feature importances should align with economic rationale.

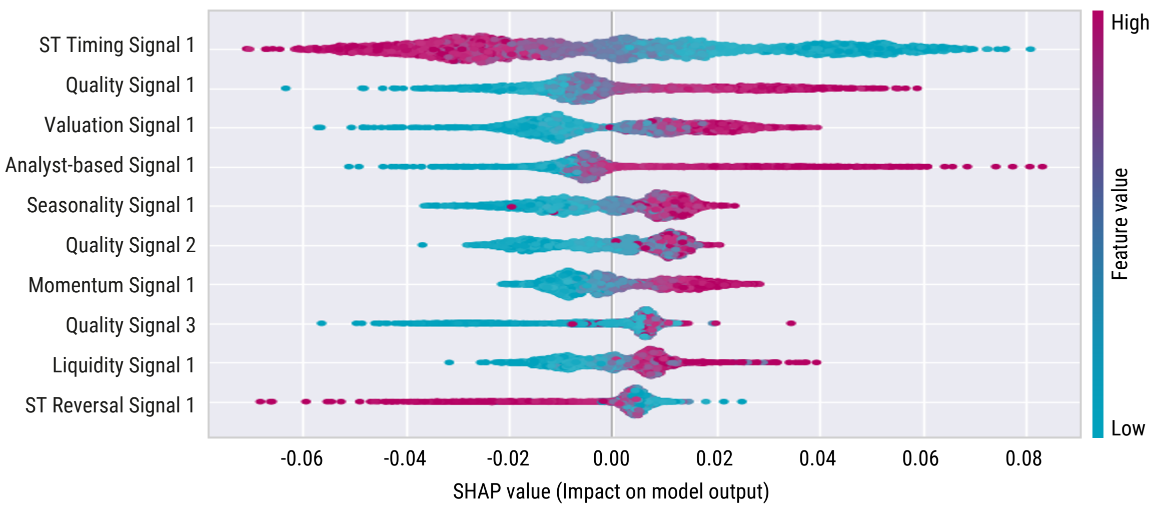
Figure 3 | Aggregate Shapley feature importance over time for 1-month-ahead return predictions



Source: Robeco. The figure shows the aggregated absolute Shapley feature importance over time for an machine learning model predicting one-month-ahead returns. The summed values are rescaled to 100% at each point in time. We display the 12 most important features, which are ranked in order of average importance, from bottom to top. The model is evaluated over the period of December 1995 to December 2024, using constituents of the MSCI World Index.

Shapley values also provide insight into how the individual variables in the model relate to final model predictions. For example, is a high value for feature X a positive or a negative predictor? For such an analysis, a Shapley summary plot is helpful. An example of such a plot is shown in Figure 4. The colors indicate the value of each of the features, and the x-axis indicates the impact on model predictions. In this case, a low value of Short-Term (ST) Timing Signal 1 (blue) is associated with high predictions (on the right side of the x-axis). Conversely, for Valuation Signal 1, a high value (red) has a positive impact on predictions.

Figure 4 | Shapley summary plot for one-month-ahead return predictions



Source: Robeco. The figure shows Shapley values for a machine learning model predicting one-month-ahead returns. Shapley values are shown on the x-axis, with a Shapley value above 0 indicating that a feature has a positive impact on model predictions. The red dots indicate high values for the respective feature and blue dots indicate low values.

ROBECO

The skewness and density of the dots in Figure 4 indicate the empirical distribution of the Shapley values. If the dots are not symmetrically distributed around the y-axis but have larger left or right tails, this indicates a nonlinear relationship between the feature and the predictions. For example, ST Reversal Signal 1 has a larger left tail. We sometimes observe that red dots and blue dots are mixed, for example, for Quality Signal 3. This implies that interaction effects are present: the impact of a feature on predictions not only depends on the value of the feature but also on the value of other features.

We can further zoom in on individual features using dependence plots, as shown in Figure 5. These dependence plots visualize how predictions relate to the values of individual features. In the plot on the left, we show that Feature X has a nonlinear relationship with predictions: the dots do not display a linear pattern but instead resemble an inverted U-shape.

A nonlinear relationship between Feature Y and the predictions is also illustrated on the right-hand side of Figure 5, which shows a 'hockey stick' pattern. Here, using a color scale, the values of a secondary Feature Z have also been indicated. The color pattern also reveals an interaction effect. We observe that the curved relationship between Feature Y and predictions is much more pronounced for low values of Feature Z (blue dots) than for high values (red dots). Thus, the relationship between Feature Y and predictions is crucially dependent on Feature Z as well.

Figure 5 | Shapley dependence plots for one-month-ahead return predictions

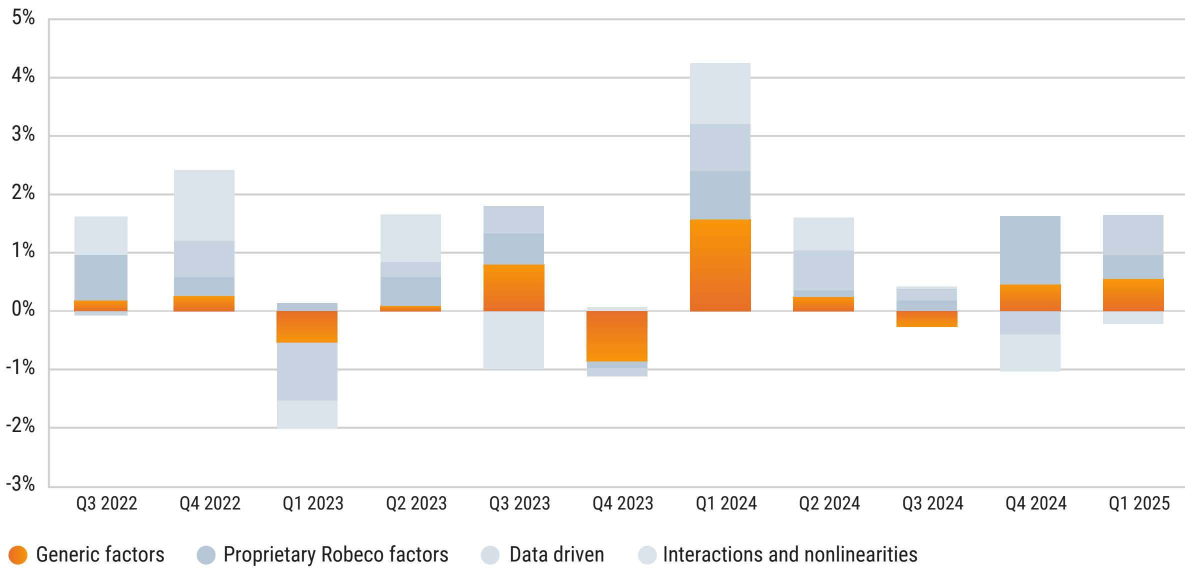
Source: Robeco. The figure shows two Shapley dependence plots for a machine learning model predicting one-month-ahead returns. Shapley values are shown on the y-axis, with a Shapley value above 0 indicating that a feature has a positive impact on model predictions. The left plot shows the dependence of the model's predictions on a single feature. The right plot enhances this by using a color scale to represent the values of secondary features, providing additional context on how these secondary features interact with the primary feature.

Attributing machine learning portfolio returns

It is crucial to understand the relationship between the input features and the resulting ML predictions, but it is equally important to understand the returns of associated investment strategies. For instance, if an investment strategy is based on ML predictions but these exposures are not being translated into the actual portfolio due to constraints, some features might have a large impact on predictions but a relatively small impact on realized strategy returns, or vice versa. To enable attributing the returns of ML-based investment strategies, we have developed a proprietary ML return attribution framework. Using this framework, we can break down the (excess) returns of ML strategies into returns stemming from different components.

In Figure 6, we show the attribution of realized ML portfolio returns, showing each of the below-listed four components of ML returns. In this example, we use monthly returns, but the same methodology can be applied to weekly or daily returns.

Figure 6 | Attribution of machine learning portfolio returns since live inception



Source: Robeco. This figure shows returns from a hypothetical top-minus-bottom self-financing portfolio which is long the top 20% stocks and short the bottom 20% stocks. Sorts are made using live, point-in-time machine learning predictions. Returns are estimated on daily data. The universe are the constituents of the MSCI World Index. The sample period is September 2022 to March 2025. The performance shown is hypothetical, based on theoretical models and assumptions, and is no guarantee of future results.

The Robeco machine learning return attribution framework

The four chosen components of our ML return attribution framework are:

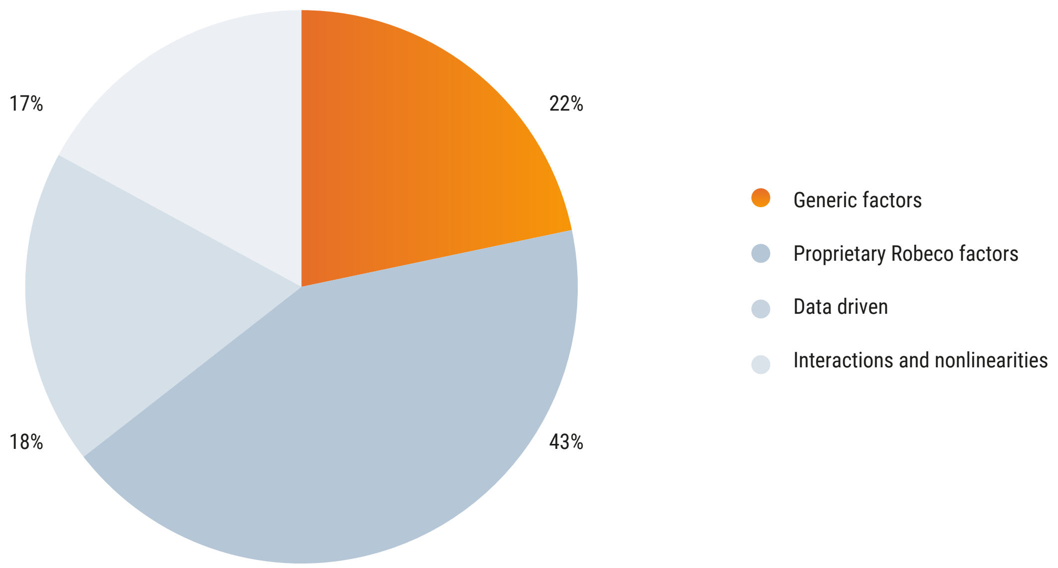
- **Generic factor definitions**: Performance stemming from exposure to generic factors (commonly used across the industry) with fixed weight
- **Proprietary Robeco factors**: Performance stemming from exposure to unique factors developed by Robeco, with fixed weight
- **Data-drivenness**: Performance stemming from ML models' data-driven factor weights can also vary over time
- **Interactions and nonlinear feature combinations**: Performance stemming from ML models' ability to capture complex relationship

These components are for illustration purposes and can be chosen flexibly based on the desired level of granularity.

Again, it is important to be able to link the analysis's outcomes to observable market developments. For example, we see a positive return from all components in the first quarter of 2024, with generic factors also making a sizeable positive contribution. Indeed, when looking at the performance of generic equity factors, they had positive returns in this quarter (most notably momentum).

We can also aggregate the attributed returns over time, as shown in Figure 7. On average, all four components had a positive contribution. About 20% of the total performance came from generic factors. Just under 20% of performance stemmed from the data-driven aspect of the ML model, and an approximately similar share was thanks to capturing nonlinear relationships and interaction effects. The remaining part is attributed to Robeco's proprietary equity factors. Note that compared to the traditional linearly combined models, the data-driven component and the nonlinearities/interactions component are what ML algorithms add to the portfolio return. Also note that this attribution over a relatively short period of time is not necessarily representative of the long-term share of each of these components.

Figure 7 | Aggregate attribution of the returns of a machine learning portfolio since live inception



Source: Robeco. This figure shows returns from a hypothetical top-minus-bottom self-financing portfolio which is long the top 20% stocks and short the bottom 20% stocks. Sorts are made using live, point-in-time ML predictions. Returns are estimated on daily data. The universe are the constituents of the MSCI World Index. The sample period is September 2022 to March 2025. The performance shown is hypothetical, based on theoretical models and assumptions, and is no guarantee of future results.

Conclusion

As machine learning techniques gain popularity in quantitative investing, the need to understand and interpret these more sophisticated models increases. To this end, we use a combination of publicly available and proprietary techniques. We believe it is essential to always be able to link back predictions and portfolio outcomes to observable characteristics, such that we can verify whether predictions are consistent with economic rationale.

We have also developed proprietary techniques to attribute the returns of machine learning investment strategies to the various underlying components. These techniques allow us to determine which share of performance can be attributed to, for example, nonlinearities and interactions, data-drivenness, Robeco proprietary factors, and generic equity factors. The enhanced transparency and interpretation from these new tools contribute to the successful use of machine learning techniques in practice. In summary, done properly, machine learning approaches to investment need not constitute an opaque 'black box' but a clearly understood 'glass box' instead.

References

Blitz, D., Hanauer, M.X., Hoogteijling, T. and Howard, C., 2023. The term structure of machine learning alpha. *Journal of Financial Data Science*, 5(4), 40-65.

Blitz, D., Hoogteijling, T., Lohre, H. and Messow, P., 2023. How can machine learning advance quantitative asset management? *Journal of Portfolio Management*, 49(9), 78-95.

Chen, M., Honarvar, I. and Lohre, H., 2023. The current state of AI for investment management. VBA Journaal, 155, 30-36.

Chen, M. and Zhou, W., 2023. Machine learning and data science applications in investments, in Cao, L. (ed.). *Handbook of AI and Big Data Applications in Investments*. CFA Institute Research Foundation, 2-19.

Chen, M., Hanauer, M., and Kalsbach, T., 2024. Design choices, machine learning, and the cross-section of stock returns. SSRN Working Paper No. 5031755.

Gu, S., Kelly, B. and Xiu, D., 2020. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223-2273.

Hanauer, M.X. and Kalsbach, T., 2023. Machine learning and the cross-section of emerging market stock returns. *Emerging Markets Review*, 55, 101022.

Shapley, L. (1953). *A Value for n-Person Games*. In: Kuhn, H. and Tucker, A., Eds., Contributions to the theory of games II, Princeton University Press, Princeton, 307-317.

Tobek, O. and Hronec, M., 2021. Does it pay to follow anomalies research? Machine learning approach with international evidence. *Journal of Financial Markets*, 56, 100588.

Important information – Capital at risk

This information refers only to general information about Robeco Holding B.V. and/or its related, affiliated and subsidiary companies, ("Robeco"), Robeco's approach, strategies and capabilities. This is a marketing communication solely intended for professional investors, defined as investors qualifying as professional clients, who have requested to be treated as professional clients or who are authorized to receive such information under any applicable laws. Unless otherwise stated, the data and information reported is sourced from Robeco, is, to the best knowledge of Robeco, accurate at the time of publication and comes without any warranties of any kind. Any opinion expressed is solely Robeco's opinion, it is not a factual statement, and is subject to change, and in no way constitutes investment advice. This document is intended only to provide an overview of Robeco's approach and strategies. It is not a substitute for a prospectus or any other legal document concerning any specific financial instrument. The data, information, and opinions contained herein do not constitute and, under no circumstances, may be construed as an offer or an invitation or a recommendation to make investments or divestments or a solicitation to buy, sell, or subscribe for financial instruments or as financial, legal, tax, or investment research advice or as an invitation or to make any other use of it. All rights relating to the information in this document are and will remain the property of Robeco. This material may not be copied or used with the public. No part of this document may be reproduced, or published in any form or by any means without Robeco's prior written permission. Robeco Institutional Asset Management B.V. has a license as manager of UCITS and AIFs of the Netherlands Authority for the Financial Markets in Amsterdam.

UNITED KINGDOM

This information is provided by Robeco Institutional Asset Management UK Limited, 30 Fenchurch Street, Part Level 8, London EC3M 3BD, registered in England no.15362605. Robeco Institutional Asset Management UK Limited is authorised and regulated by the Financial Conduct Authority (FCA – Reference No: 1007814).

SWITZERLAND

Robeco Switzerland Ltd is licensed by the Swiss Financial Market Supervisory Authority FINMA as a manager of collective assets.

AUSTRALIA

This document is distributed in Australia by Robeco Hong Kong Limited (ARBN 156 512 659) ('Robeco') which is exempt from the requirement to hold an Australian financial services license under the Corporations Act 2001 (Cth) pursuant to ASIC Class Order 03/1103. Robeco is regulated by the Securities and Futures Commission under the laws of Hong Kong and those laws may differ from Australian laws.

HONG KONG

Issued by Robeco Hong Kong Limited, licensed and regulated by Securities and Futures Commission of Hong Kong. The contents of this document have not been reviewed by the Securities and Futures Commission Hong Kong. Investment involves risks. This information does not constitute an offer to sell, a solicitation of an offer to buy, or a recommendation for any security.

SINGAPORE

This information is for informational purposes only and should not be construed as an offer to sell or an invitation to buy any securities or products, nor as investment advice or recommendation. The contents of this document have not been reviewed by the Monetary Authority of Singapore ("MAS").

Robeco Singapore Private Limited holds a capital markets services licence for fund management issued by the MAS and is subject to certain clientele restrictions under such licence. An investment will involve a high degree of risk, and you should consider carefully whether an investment is suitable for you.

US

This information is for informational purposes only and should not be construed as an offer to sell or an invitation to buy any securities or products, nor should it be viewed as investment advice or a recommendation to buy or sell any financial instrument or instruments.

Investment management services are offered to U.S. institutional clients by Robeco Institutional Asset Management U.S. Inc. ("RIAM US"), an investment adviser registered with the US Securities and Exchange Commission ("SEC"). Such registration should not be viewed as an endorsement or approval of RIAM US by the SEC. RIAM US is a wholly owned subsidiary of ORIX Corporation Europe N.V. ("ORIX"), a Dutch investment management firm headquartered in Rotterdam, the Netherlands. RIAM US is located at 230 Park Avenue, 33rd floor, New York, NY 10169.