Machine Learning Mutual Fund Flows

Jürg Fausch[1], Moreno Frigg[1,3], Stefan Ruenzi[2], Florian Weigert[3]

American Finance Association (AFA) Annual Meeting 2026
Philadelphia, USA

January 5, 2026

[1]Lucerne University of Applied Sciences & Arts (HSLU)

[2]University of Mannheim

[3]University of Neuchâtel

Motivation for the study

- Fund flows are volatile and hard to predict $\Rightarrow$ liquidity management difficult (Edelen, 1999)
- Many variables (past performance, fund characteristics,...) influencing flows have been documented (e.g Sirri and Tufano, 1998)
- Traditional models fail to capture interactions & nonlinearities
- Ideal setting for ML to improve accuracy and offer new insights (interpretable ML)

Data - Sample

Main data sources:

- CRSP Survivor-Bias-Free US Mutual Fund database
- MorningstarDirect

Sample (monthly data):

- Actively managed US equity fund share classes from 1991 to 2023
- Filters to avoid incubation bias (Elton et al., 2001; Evans, 2010; Doshi et al., 2015; Kaniel et al., 2023)

Merged sample consists of **9,721** unique share classes (retail and institutional).

Data - Preprocessing

1. **Target variable:** Our target variable is defined as fund flows in $t+1$

2. **Definition of fund flows:**

$$flow_{i,t} = \frac{TNA_{i,t} - TNA_{i,t-1}\left(1 + R_{i,t}\right)}{TNA_{i,t-1}}, \tag{1}$$

3. **64 predictor variables:**
   - Performance measures (fund returns, different contemporaneous and lagged (past 3 to 18 months) factor model alphas, Sharpe ratio, MS rating,...)
   - Fund characteristics (size, fees, turnover, age,...)
   - Flow-related variables to account for fund flow persistence (past mean flows over a six and 12-month horizon,...)
   - Macroeconomic variables (market returns, market volatility, interest rates, term spread, various risk aversion and uncertainty indices,...)

4. **No data imputation (drop NA)**

Prediction framework - ML methods and benchmark OLS

Applied algorithms:

- OLS (benchmark model), allowing for non-linear impact of past performance (Chevalier and Ellison, 1997; Sirri and Tufano, 1998)
- Elastic net
- Decision tree
- Random forest
- Histogram based gradient boosting
- Feedforward neural networks
  - NN-1: 1-3 hidden layers and 2-32 neurons per hidden layer
  - NN-2: 3-10 hidden layers and 32 to 1024 neurons per hidden layer.

Prediction framework - Model Tuning & Calibration

Model tuning/calibration:

- Initial training + validation sample: 01/1991 - 11/1999 (expanding)
- Goal: Find hyperparameters associated with the best model fit
- k-fold cross validation ($k = 5$); random validation set (for NNs)
- Forecast accuracy based on *MSFE*
- Hyperparameters updated bi-annually

Applying the REFORMS Checklist

- Many ML applications suffer from various pitfalls and problems: Modeling choices, parameter tuning, and potential contamination of results due to forward-looking bias
- Kapoor et al. (2024) develop a consensus-based **Re**commendations **for** **M**achine-learning-based **S**cience (REFORMS) checklist
- We apply their suggested procedure in a finance context.

Prediction framework - Evaluation of OOS predictive ability

Prediction:

Optimally calibrated model used to compute flow forecast for $t+1$

Evaluation of predictive power:

- R^2_{OOS} (Welch and Goyal, 2008; Gu et al., 2020) to evaluate out-of-sample (OOS) performance, using historical mean flows as benchmark.

- Statistical significance based on the Clark and West (2007) statistic:

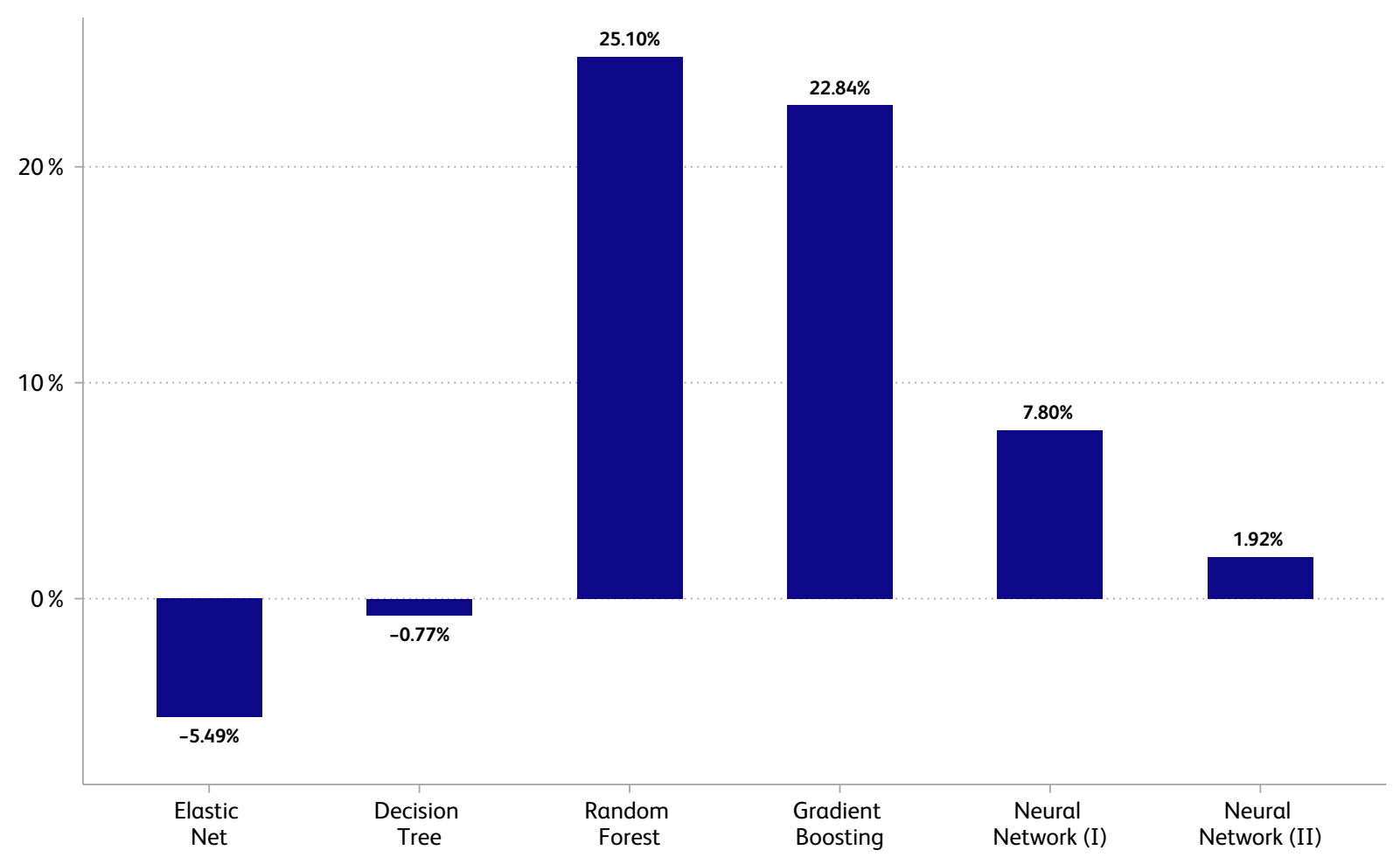
$$H_0 : R^2_{OOS} = 0 \text{ against } H_1 : R^2_{OOS} \neq 0.$$

- Modified Diebold-Mariano-West (DMW) tests (Diebold and Mariano, 1995; West, 1996; Harvey et al., 1997) to compare forecast accuracy of models.

▸ Statistical tests details

Results: Out-of-sample predictability - OOS R^2 i

Figure 1: Relative improvement in out-of-sample R^2 vs. OLS (full sample)



Results: Out-of-sample predictability - OOS R^2 ii

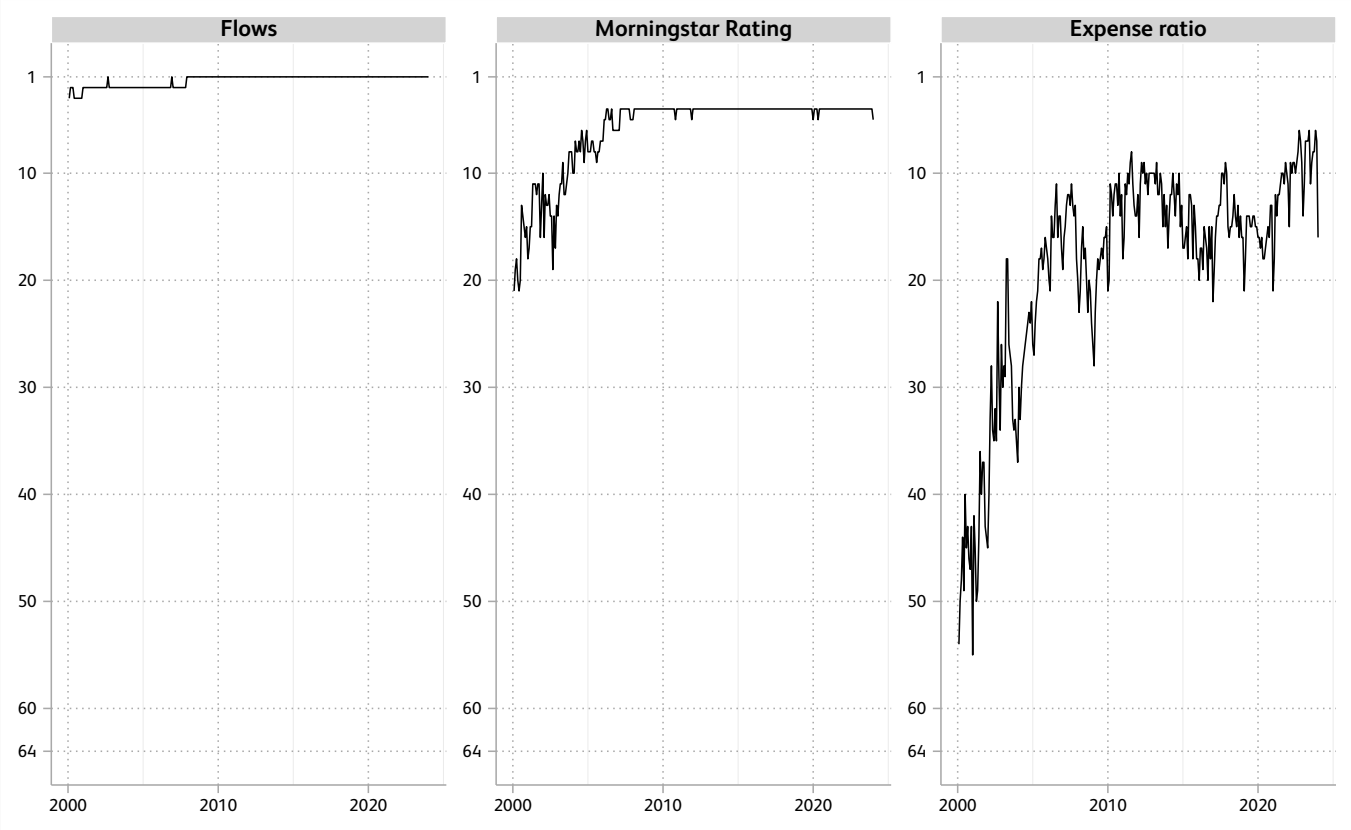
	Out-of-sample R^2		
	Full sample	Top 10%	Bottom 10%
OLS - FF4	0.1821***	0.0131***	0.1941***
Elastic net	0.1721***	-0.0186***	0.1760***
Decision tree	0.1807***	0.0375***	0.1739***
Random forest	0.2278***	0.1088***	0.2010***
Gradient boosting	0.2237***	0.1025***	0.1980***
Neural network (I)	0.1963***	0.0407***	0.2005***
Neural network (II)	0.1918***	0.0268***	0.2002***
Observations	1,106,802	110,686	110,686

Interpretable Machine Leaning using SHAP values

- Understanding the reasoning behind predictions is critical (interpretable ML).
- Model-agnostic SHAP values (Lundberg and Lee, 2017; Lundberg et al., 2020) allow to assess feature importance for prediction.
- The overall importance of each characteristic is based on mean absolute SHAP values, representing the average strength of each predictor's impact on the model prediction.
- Most important predictors (based on random forest):
  lagged flows, MS rating, size, CAPM alpha, flow volatility, expense ratio,...

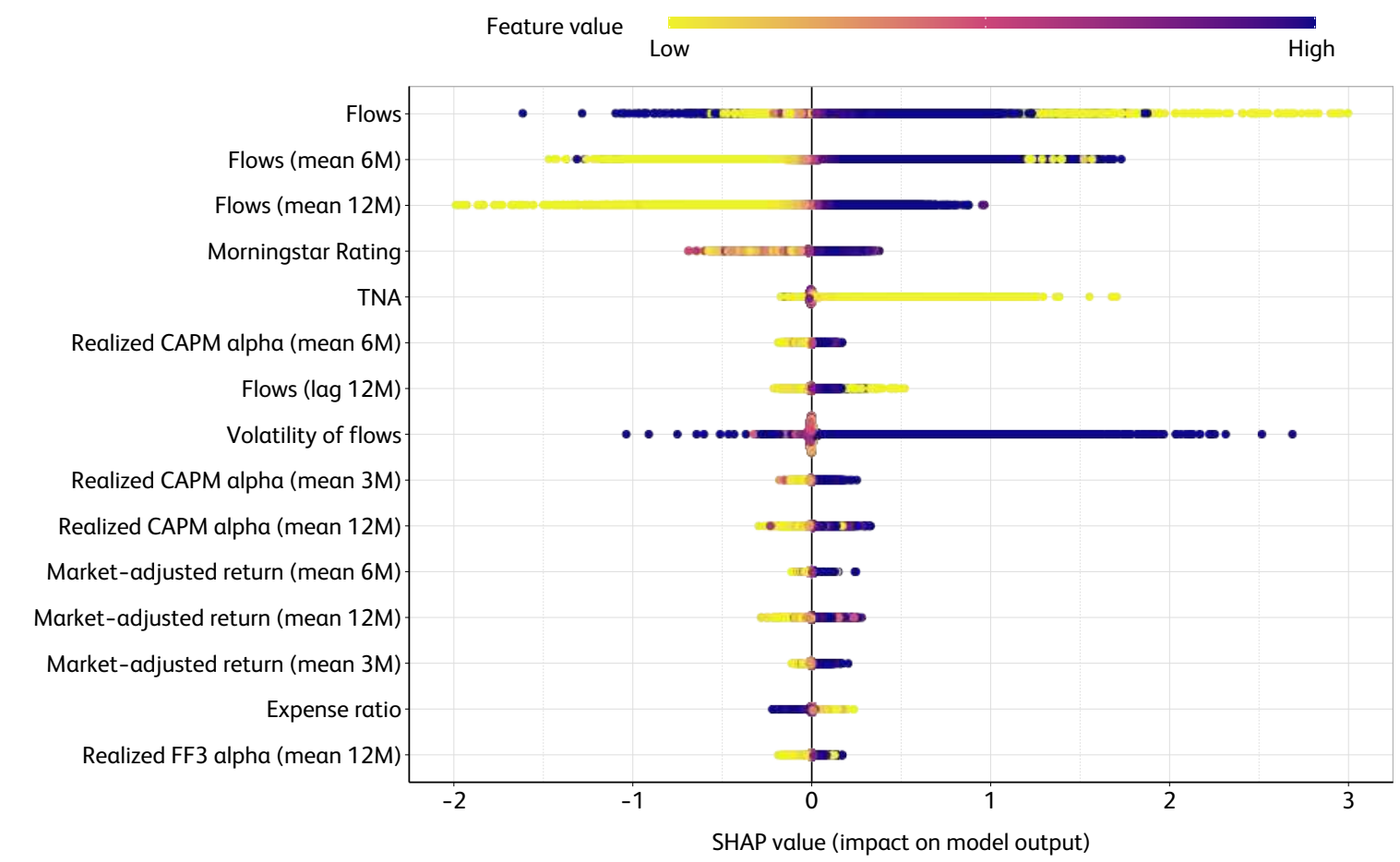
Results: Importance over time

Figure 2: Evolution of characteristic importance for random forest over time



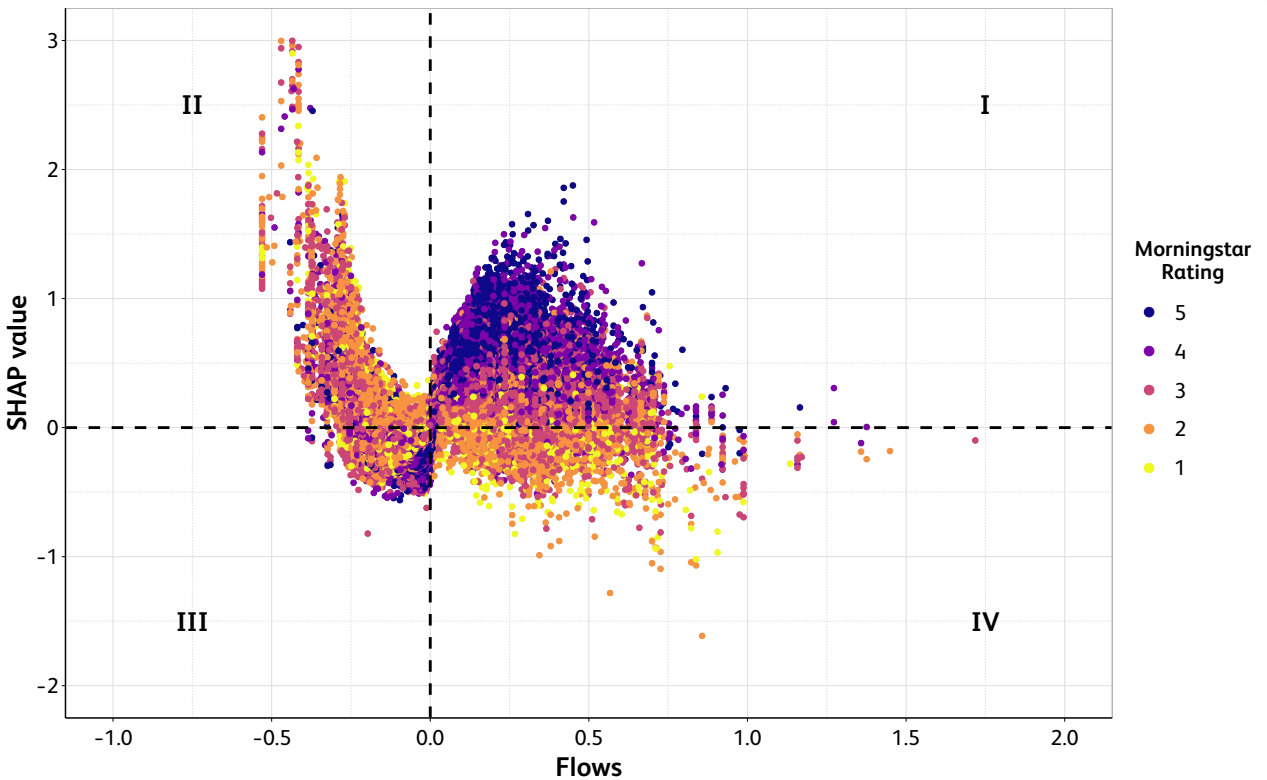
Results: Directional impact

Figure 3: Magnitude and directional impact of the characteristics for random forest



Results: Interaction and Non-Linearties

Figure 4: Interaction effect of past fund flows and Morningstar rating



- High past flows & high (low) MS ratings lead to high (low) flow predictions.
- Suggests that “new money” is performance sensitive.

Results: Predicted Fund Flows and Performance

Table 1: Monthly OOS α (in %), long-short 1th percentile portfolios

	CAPM	FF3	FF4	FF5
OLS	0.33	0.30	0.21	0.21
	(0.21)	(0.21)	(0.18)	(0.18)
Random Forest	0.33**	0.30**	0.25*	0.27**
	(0.12)	(0.11)	(0.10)	(0.09)
Gradient Boosting	0.38***	0.36***	0.32**	0.33***
	(0.14)	(0.13)	(0.12)	(0.11)

- Annual alpha of about 3.5%.
- Consistent with smart-money effect/adverse effects of large outflows.

Prediction Accuracy and Performance

- Hypothesis: Funds, where flow prediction accuracy based on basic models (historical mean) is worse than those from ML models suffer from difficulties in liquidity management and eventually also perform worse.

- Results show that these funds significantly underperform by an α of -0.53% to -0.73% per year

Conclusion - ML unlocks the why

1. Nonlinear machine learning methods significantly outperform linear models in terms of out-of-sample R^2.
2. Interpretable ML helps to decode investor behavior:
   - Past flows and the Morningstar (MS) rating are identified as the most important predictors (SHAP values).
   - Importance of the predictors is time-varying. The relevance of the MS rating and the expense ratio has increased over time.
3. Interaction of past flows with MS rating has strong impact on flows, suggesting "new money" is performance sensitive.
4. ML-based flow predictions have performance implications and generate economically and statistically highly significant alphas based on long-short portfolios.
5. The lower the forecast accuracy for a fund, the worse is its performance.

Thank you very much!

Fausch/Frigg/Ruenzi/Weigert (2025):
Machine Learning Mutual Fund Flows, Working Paper

References i

Bergstra, J., R. Bardenet, Y. Bengio, and B. Kégl (2011): "Algorithms for hyper-parameter optimization," *Advances in neural information processing systems*, 24.

Chevalier, J. and G. Ellison (1997): "Risk taking by mutual funds as a response to incentives," *Journal of Political Economy*, 105, 1167–1200.

Clark, T. E. and K. D. West (2007): "Approximately normal tests for equal predictive accuracy in nested models," *Journal of econometrics*, 138, 291–311.

Diebold, F. X. and R. S. Mariano (1995): "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

Doshi, H., R. Elkamhi, and M. Simutin (2015): "Managerial activeness and mutual fund performance," *The Review of Asset Pricing Studies*, 5, 156–184.

Edelen, R. M. (1999): "Investor flows and the assessed performance of open-end mutual funds," *Journal of Financial Economics*, 53, 439–466.

Elton, E. J., M. J. Gruber, and C. R. Blake (2001): "A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund databases," *The Journal of Finance*, 56, 2415–2430.

Evans, R. B. (2010): "Mutual fund incubation," *The Journal of Finance*, 65, 1581–1611.

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

References ii

Harvey, D., S. Leybourne, and P. Newbold (1997): "Testing the equality of prediction mean squared errors," *International Journal of forecasting*, 13, 281–291.

Kaniel, R., Z. Lin, M. Pelger, and S. Van Nieuwerburgh (2023): "Machine-learning the skill of mutual fund managers," *Journal of Financial Economics*, 150, 94–138.

Kapoor, S., E. M. Cantrell, K. Peng, T. H. Pham, C. A. Bail, O. E. Gundersen, J. M. Hofman, J. Hullman, M. A. Lones, M. M. Malik, et al. (2024): "REFORMS: Consensus-based Recommendations for Machine-learning-based Science," *Science Advances*, 10, eadk3452.

Li, L., K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar (2017): "Hyperband: A novel bandit-based approach to hyperparameter optimization," *The Journal of Machine Learning Research*, 18, 6765–6816.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee (2020): "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, 2, 56–67.

Lundberg, S. M. and S.-I. Lee (2017): "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 30.

References iii

Newey, W. and K. West (1987): "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–08.

Sirri, E. R. and P. Tufano (1998): "Costly search and mutual fund flows," *The Journal of Finance*, 53, 1589–1622.

Welch, I. and A. Goyal (2008): "A comprehensive look at the empirical performance of equity premium prediction," *The Review of Financial Studies*, 21, 1455–1508.

West, K. D. (1996): "Asymptotic inference about predictive ability," *Econometrica*, 1067–1084.

Appendix: Details of model tuning

- Initial training + validation sample: January 1989 to November 1999 (expanding)
- Goal: Find hyperparameters of ML methods associated with the best model fit.
- k-fold cross validation ($k = 5$); random validation set (for neural networks)
- Tree-structured Parzen estimator approach (from *Optuna*) suggested by Bergstra et al. (2011)
- Additionally, we use Hyperband (Li et al., 2017) as a pruning mechanism for neural networks
- Metric of interest to evaluate forecast accuracy: *MSFE*

◂ Go back

Appendix: Modified Diebold-Mariano-West Statistics

$$\hat{d}_t^{(1,2)} = \left(\hat{e}_{i,t+h}^{(1)}\right)^2 - \left(\hat{e}_{i,t+h}^{(2)}\right)^2 \tag{A.1}$$

The DMW statistic to test the null hypothesis of equal predictive accuracy, $H_0 : E\left(\hat{d}_t^{(1,2)}\right) = 0$, against the two-sided alternative is obtained as the t-statistic of a regression with intercept only, $\hat{d}_t^{(1,2)} = \mu + \varepsilon_t$:

$$DMW^{(1,2)} = \frac{\hat{\mu}^{(1,2)}}{\hat{\sigma}_{\hat{\mu}}^{(1,2)}} \tag{A.2}$$

where $\hat{\mu}^{(1,2)}$ denotes the estimated coefficient (time series average of $\hat{d}_t^{(1,2)}$) and $\hat{\sigma}_{\hat{\mu}}^{(1,2)}$ is the corresponding Newey and West (1987) HAC standard error. To obtain improved small-sample properties we follow Harvey et al. (1997) and make a bias correction to the DMW statistic in (A.2). The corrected test statistic is obtained as

$$HLN - DMW^{(1,2)} = \sqrt{\frac{T + 1 - 2h + T^{-1}h(h-1)}{T}} DMW^{(1,2)}. \tag{A.3}$$

which is compared to the critical values of a Student t-distribution.

◂ Go back