Machine Learning for Finance (FIN 570)
Machine Learning in Finance Research

Instructor: Jaehyuk Choi

Peking University HSBC Business School, Shenzhen, China

2023-24 Module 3 (Spring 2024)

NN in Finance Research 1: Stock return

- Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. The Review of Financial Studies, 33(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009
- A comparative (and educational) analysis of ML methods for predicting cross sectional stock returns. Trees and NNs performs best, identifying momentum, liquidity, and volatility as the key predictors.
- Monthly US stock returns over 60 years, 94 stock characteristics, 74 industry dummies, 8 macro economic predictors.
- Performance evaluation: out-of-sample R^2 gain:

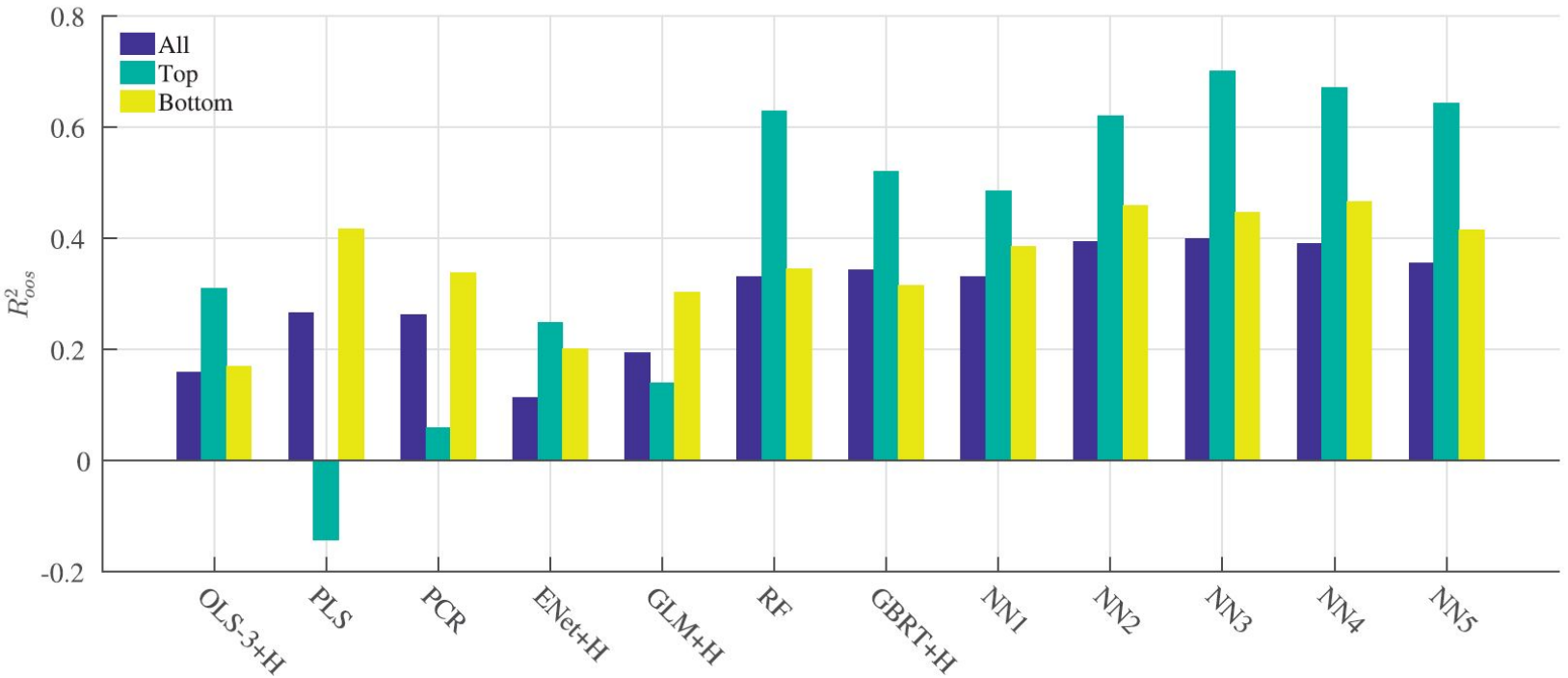
$$R^2_{\text{OOS}} = 1 - \frac{\sum_{i,t}(r_{i,t} - \hat{r}_{i,t})^2}{\sum_{i,t}(r_{i,t} - \bar{r})^2} \quad \text{for stock } i \text{ and test period } t, \text{ and } \bar{r} = 0.$$

- NN specifications: up to 5 fully-connected hidden layers (32, 16, 8, 4, 2). ReLU activation. Early stopping and L_1 regularization.
- In early stopping, the SGD iteration stops when test accuracy is not improving (although training accuracy is still improving).

- Prediction performance in Gu, Kelly, & Xiu. (2021):

Table 1
Monthly out-of-sample stock-level prediction performance (percentage R^2_{OOS})

	OLS +H	OLS-3 +H	PLS	PCR	ENet +H	GLM +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
All	−3.46	0.16	0.27	0.26	0.11	0.19	0.33	0.34	0.33	0.39	0.40	0.39	0.36
Top 1,000	−11.28	0.31	−0.14	0.06	0.25	0.14	0.63	0.52	0.49	0.62	0.70	0.67	0.64
Bottom 1,000	−1.30	0.17	0.42	0.34	0.20	0.30	0.35	0.32	0.38	0.46	0.45	0.47	0.42

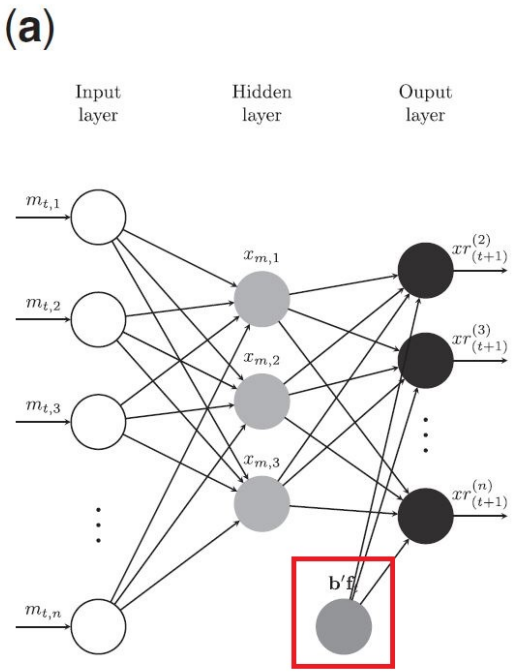


NN in Finance Research 2: Bond return

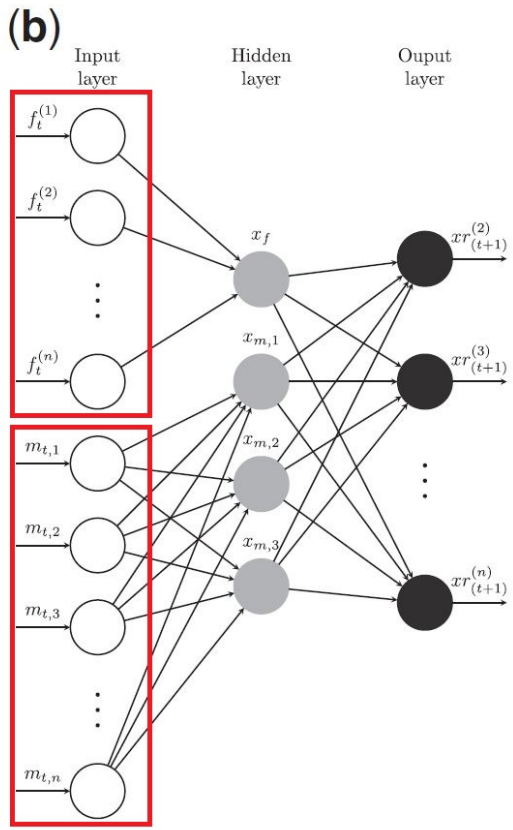
- Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond Risk Premiums with Machine Learning. The Review of Financial Studies. https://doi.org/10.1093/rfs/hhaa062
- Extreme trees and NNs predict the bond return well.
- In NN, yields plus macroeconomic factors predicts better than yields only, against the *spanning hypothesis* that the information relevant to the bond excess return is wholly contained in the current yield curve.
- Data: monthly yield curve up to 10y, 128 monthly macroeconomic and financial variables.
- Performance evaluation: out-of-sample R^2 gain:

$$R^2_{\text{OOS}} = 1 - \frac{\sum_t (xr_t^{(n)} - \hat{xr}_t^{(n)})^2}{\sum_t (xr_t^{(n)} - \bar{xr}^{(n)})^2} \quad \text{for test period } t \text{ and maturity } n.$$

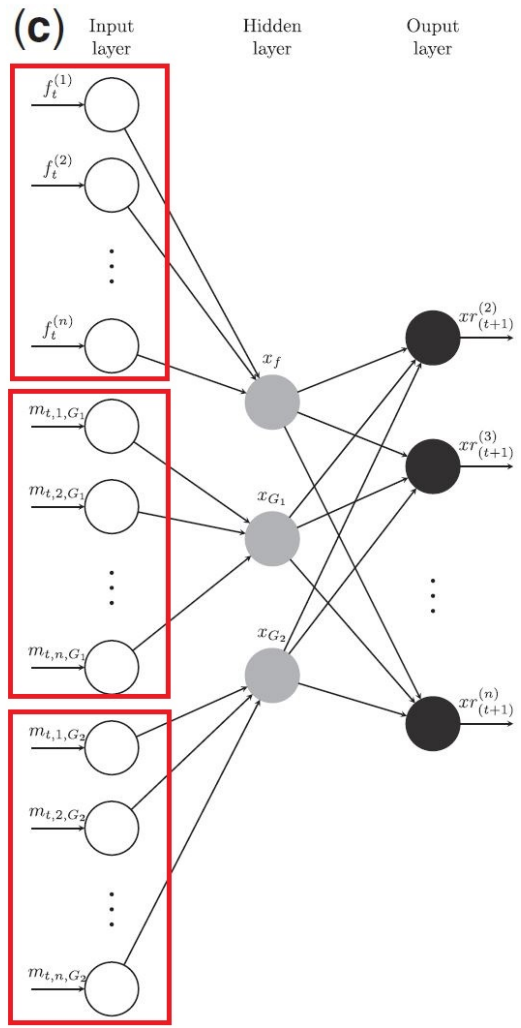
- Various NN specifications used by Bianchi et al. (2021): **fwd rates only**, **(a) hybrid**, **(b) macro + fwd rate**, **(c) group-ensemble**.



Hybrid network with fwd rates direct

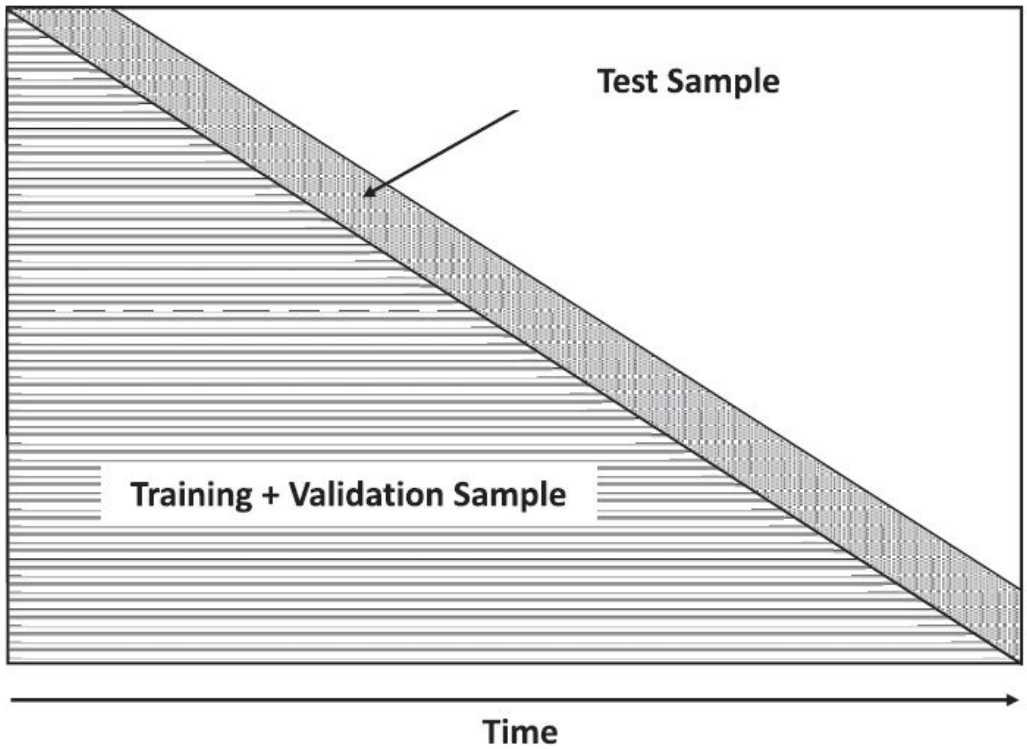


Macro network + fwd rate net



Group-ensemble network

- In financial ML, randomizing time series cause look-ahead bias. We preserve the time order in training-test split.
- Training vs test period used in Bianchi et al. (2021)



NN in Finance Research 3: Autoencoder

Gu, S., Kelly, B., & Xiu, D. (2020). Autoencoder asset pricing models. Journal of Econometrics. https://doi.org/10.1016/j.jeconom.2020.07.009

- Autoencoder is a special type of NN with the same input and output. So the network is trained to reconstruct the same input as much as possible. See **PML** Ch. 17.
- It is decomposed into encoder and decoder.
- The low-dimensional *bottleneck* is the latent vector. It is understood as the data compression or nonlinear PCA.
- In asset pricing, the encoded data is understood as the asset pricing "factors".