Paper Replication: Empirical Asset Pricing via Machine Learning

Runhao SHI

Hong Kong University of Science and Technology

https://www.bilibili.com/video/BV1nY4y1k7xN/

Table of Contents

1 Introduction

2 Methodology

3 Experiments

4 Conclusion

Introduction

- [Gu et al., 2020] perform a comparative analysis of machine learning methods in measuring asset risk premiums (risk premiums refer to the conditional expected stock returns in excess of the risk-free rate).
- We replicate the machine learning methods used in [Gu et al., 2020].
- Specifically, we replicate 7 methods mentioned in this paper, including ordinary least squares (OLS), penalized linear with elastic net penalty (ENet), principal components regression (PCR), partial least squares (PLS), random forests (RF), gradient boosted regression trees (GBRT), and neural networks (NN).

Table of Contents

1 Introduction

2 Methodology

3 Experiments

4 Conclusion

Simple linear (OLS)

- The simple linear model using ordinary least squares (OLS) estimation could be written as the following optimization problem

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} H(r_{i,t+1} - g(\mathbf{z}^{(i;t)}, \boldsymbol{\theta}); \xi), \tag{1}$$

where

$$g(\mathbf{z}^{(i,t)}; \boldsymbol{\theta}) = \boldsymbol{\theta}^{\top}\mathbf{z}^{(i,t)}, \tag{2}$$

and

$$H(x; \xi) = \begin{cases} x^2, & \text{if } |x| \leq \xi \\ 2\xi|x| - \xi^2, & \text{if } |x| > \xi. \end{cases} \tag{3}$$

Penalized linear (ENet)

- The simple linear model (2) may fail in the presence of many predictors. In order to enforce the sparsity of the final model, [Gu et al., 2020] append elastic net penalty to (1) as the following

$$\phi(\boldsymbol{\theta}; \lambda, \rho) = \lambda(1-\rho)\sum_{j=1}^{P}|\theta_j| + \frac{1}{2}\lambda\rho\sum_{j=1}^{P}\theta_j^2. \tag{4}$$

Principal components regression (PCR) and partial least squares (PLS)

- Since the penalized linear model may achieve sub-optimal forecasts when predictors are highly correlated, [Gu et al., 2020] introduce two dimension reduction techniques. The linear regression could be represented as

$$\mathbf{R} = \mathbf{Z}\boldsymbol{\theta}, \tag{5}$$

where $\mathbf{R} \in \mathbb{R}^{NT}$ and $\mathbf{Z} \in \mathbb{R}^{NT \times P}$. Both principal components regression (PCR) and partial least squares (PLS) condense the dimension of predictors from P to K by the following

$$\mathbf{R} = (\mathbf{Z}\boldsymbol{\Omega}_K)\boldsymbol{\theta}_K, \tag{6}$$

where $\boldsymbol{\Omega}_K \in \mathbb{R}^{P \times K}$ and $\boldsymbol{\theta}_K \in \mathbb{R}^K$.

Gradient boosted regression trees (GBRT) and random forests (RF)

- To incorporate multiway interactions of predictor, [Gu et al., 2020] adopt regression trees to capture correlations among predictors. The regression tree with K leaves and depth L has the following form

$$g(\mathbf{z}^{(i,t)}; \boldsymbol{\theta}, K, L) = \sum_{k=1}^{K} \theta_k \mathbb{I}_{\mathbf{z}^{(i,t)} \in C_k(L)}, \tag{7}$$

where $C_k(L)$ denotes one of the K partitions of the data. To address the issues of overfitting, [Gu et al., 2020] apply two ensemble tree regularizers: gradient boosted regression trees (GBRT) and random forests (RF).

Neural networks (NN)

- The last method used in [Gu et al., 2020] is the traditional feed-forward neural networks (NN). The activation function used in proposed neural networks is the rectified linear unit (ReLU). The diagrams of neural network is shown in the following Figure.

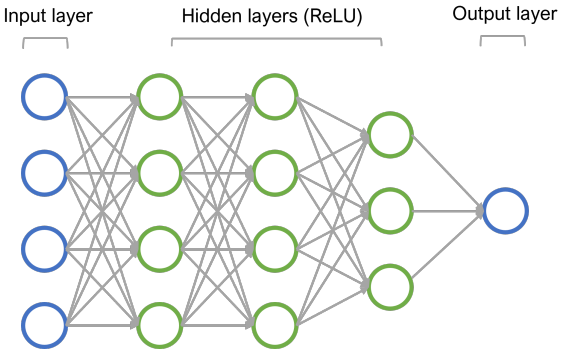

Figure: Neural networks architecture

Table of Contents

1 Introduction

2 Methodology

3 Experiments

4 Conclusion

Performance evaluation

- We use three indicators contained in [Gu et al., 2020] to evaluate machine learning methods and relative importance of predictors in estimating excess returns.
- The first one is to assess the predictive performance, which is out-of-sample R^2 as the following

$$R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in\mathcal{I}_3}(r_{i,t+1} - \hat{r}_{i,t+1})}{\sum_{(i,t)\in\mathcal{I}_3} r^2_{i,t+1}}. \tag{8}$$

- We then use Diebold and Mariano test to make pairwise comparisons of different machine learning methods. The DM statistics is $DM_{12} = \bar{d}_{12}/\hat{\delta}_{\bar{d}_{12}}$, where

$$d_{12,t+1} = \frac{1}{n_{3,t+1}} \sum_{i=1}^{n_{3,t+1}} \left((\hat{e}^{(1)}_{i,t+1}) - (\hat{e}^{(2)}_{i,t+1}) \right). \tag{9}$$

- We also compare the variable importance of different predictors, which is denoted as VI_j for the jth predictor. VI_j is the reduction of R^2 from setting all values of predictor j to zero, while holding the remaining model estimate fixed.

Data preparation and model specification

- The data we used are from Dacheng Xiu's web site and Amit Goyal's web site.
- To be specific, our dataset contains 94 stock characteristics, 8 macroeconomics predictors and 74 industry dummies corresponding to SIC codes.
- The hyper-parameters for all methods are shown in the following Table.

	OLS +H	PLS	PCR	ENet +H	RF	GBRT +H	NN
Huber loss	✓	-	-	✓	-	✓	-
#covariates	176	176	176	176	176	176	920
Others				$\rho = 0.5$ $\lambda \in (10^{-5}, 10^{-2})$	Depth= $1 \sim 6$ #Trees= 50 #Features in each split $= 8$	Depth= $1 \sim 2$ #Trees= 50 Learning rate= 0.1	L1 penalty $\lambda_1 = 10^{-3}$ Learning rate= 10^{-2} Batch Size= 10^4 Epochs= 100 Patience= 5 Ensemble= 10

Table: Hyperparameters for all methods

Out-of-sample performance

- We first show the out-of-sample R^2_{oos} of machine learning methods in the following Table and Figure.

	OLS +H	OLS-3 +H	PLS	PCR	ENet +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
All	-7.94	-0.01	-0.83	0.29	0.12	0.24	-0.1	-0.2	0.18	0.28	0.28	0.28
Top 1,000	-31.53	1.23	-4.46	1.53	0.51	1.2	-0.62	-0.68	0.79	1.73	1.8	1.77
Bottom 1,000	-5.29	-0.03	-0.4	-0.2	-0.02	-0.13	-0.19	-0.14	-0.19	-0.2	-0.2	-0.2

Table: Monthly out-of-sample stock-level prediction performance (percentage R^2_{oos}).

Out-of-sample performance

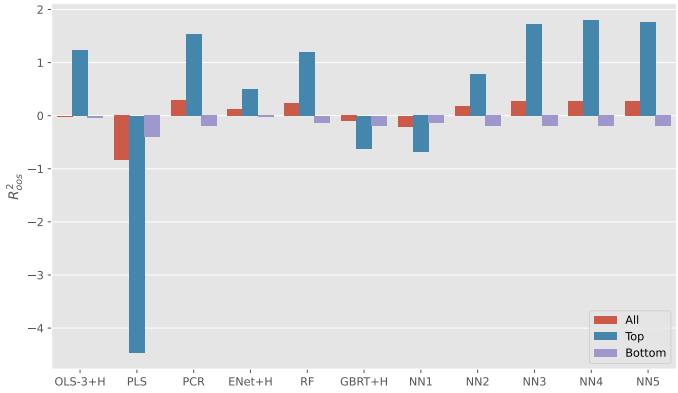

Figure: Monthly out-of-sample stock-level prediction performance (percentage R^2_{oos}).

Out-of-sample performance

- The following Table assesses the statistical significance of differences among models. Bold numbers denote significance at the 5% level or better for each individual test. We could see that for PCR has the best performance among all linear models while NN has the best performance among all non-linear models.

	OLS-3 +H	PLS	PCR	ENet +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
OLS+H	**8.08**	**9.26**	**9.09**	**8.26**	**9.31**	**8.47**	**8.73**	**8.88**	**8.99**	**9.02**	**8.95**
OLS-3+H	0	-2.67	**1.84**	**2.79**	1.20	-0.34	-0.77	1.21	**1.75**	**1.72**	**1.67**
PLS	**2.67**	0	**3.92**	**3.21**	**3.85**	**2.90**	**2.30**	**3.48**	**3.69**	**3.75**	**3.62**
PCR	-1.84	-3.92	0	-1.26	-0.20	-1.56	-3.98	-1.59	-0.25	-0.47	-0.54
ENet+H	-2.79	-3.21	1.26	0	0.67	-1.11	-1.59	0.53	1.17	1.14	1.09
RF	-1.20	-3.85	0.20	-0.67	0	-1.47	-2.3	-0.36	0.14	0.13	0.11
GBRT+H	0.34	-2.90	1.56	1.11	1.47	0	-0.41	1.07	1.44	1.48	1.42
NN1	0.77	-2.30	**3.98**	1.59	**2.30**	0.41	0	**2.94**	**3.84**	**3.96**	**3.84**
NN2	-1.21	-3.48	1.59	-0.53	0.36	-1.07	-2.94	0	1.41	1.34	1.38
NN3	-1.75	-3.69	0.25	-1.17	-0.14	-1.44	-3.84	-1.41	0	-0.1	-0.21
NN4	-1.72	-3.75	0.47	-1.14	-0.13	-1.48	-3.96	-1.34	0.1	0	-0.17
NN5	-1.67	-3.62	0.54	-1.09	-0.11	-1.42	-3.84	-1.38	0.21	0.17	0

Table: (All) Comparison of monthly out-of-sample prediction using Diebold-Marianon tests.

Comparison of variable importance

- We now investigate the relative importance of individual predictors for the performance of each model using the variable importance measures.

Figure: Variable importance by model

Comparison of variable importance

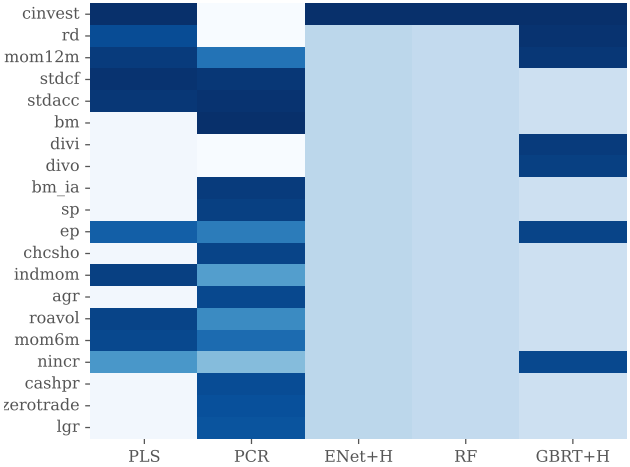

Figure: Characteristic importance.

Comparison of variable importance

- The Figure shows the R^2_{oos}-based importance measure for macro-predictors.
- We could see that dividend-price ratio (dp) and default spread (dfy) are critical predictors, and stock variance (svar) is the least informative macroeconomic predictors compared to others.

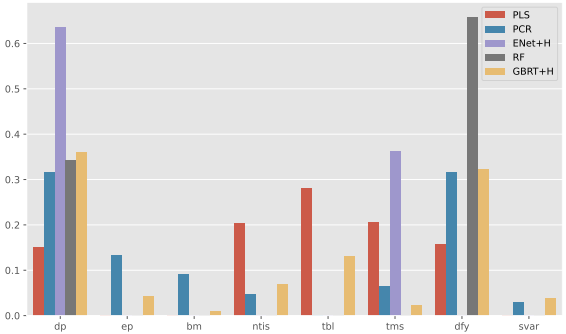

Figure: Variable importance for macroeconomic predictors.

Table of Contents

1. Introduction
2. Methodology
3. Experiments
4. Conclusion

Conclusion

- From the out-of-sample performance, we conclude that PCR has the best performance among all linear models while NN has the best performance among all non-linear models.
- As for neural networks, a three-layer might be the most proper model architecture for estimating excess returns since the improvement of four- and five-layer models is limited.
- By evaluating the relative importance of individual characteristics for the performance of each model, we find that high-importance characteristics are less noisy than low-importance characteristics.

Thanks!

References I

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