Paper Replication: Empirical Asset Pricing via Machine Learning by Gu, Kelly, and Xiu (2018)[1]

Zhen HOU, Jianda MAO, Xiaolong WANG

April 29, 2024

[1] https://youtu.be/6NaL8qs5bfM

Outline

1. Introduciton
2. Dataset and Preprocessing
3. Machine Learning Models
4. Performace Evaluation
5. Variable Importance
6. Conclusion

In this work, the excess return of an asset is modeled by:

$$\mathrm{r_{i,t+1}} = \mathbb{E}_\mathrm{t}(\mathrm{r_{i,t+1}}|\mathrm{z_{i,t}}) + \epsilon_\mathrm{i,t+1}, \tag{1}$$

where

$$\mathbb{E}_\mathrm{t}(\mathrm{r_{i,t+1}}|\mathrm{z_{i,t}}) = \mathrm{g}^\star(\mathrm{z_{i,t}}). \tag{2}$$

Notation:
Stockes $\mathrm{i} = 1, \ldots, \mathrm{N_t}$; months $\mathrm{t} = 1, \ldots, \mathrm{T}$.
$\mathrm{r_{i,t+1}}$: the excess return, serving as reponse variable.
$\mathrm{z_{i,t}}$: vector of predictor variable.
$\mathrm{g}^\star(\cdot)$: conditional expectation of $\mathrm{r_{i,t+1}}$, e.g. risk premium, given $\mathrm{z_{i,t}}$.
The main goal: to estimate $\mathrm{g}^\star(\cdot)$, a typical prediction task and attractive for machine learning methods.

Outline

Raw Dataset

- Obtained from[2].
- contains almost 30000 stocks from 1957 to 2020.
- average 6200 stocks per month.
- DATE: the end day of each month t.
- RET: lag-adjusted CRSP returns $r_{i,t+1}$.
- 94 characteristics $c_{i,t}$.
- sic2: first two digits of SIC code.

[2] https://www.dropbox.com/s/zzgjdubvv23xkfp/datashare.zip?dl=0

Cleaning and Construction

We impute missing values for each characteristic by the median of the corresponding subgroup of stocks classified by the belonging month. We include 74 industry dummies $\mathrm{s}_{\mathrm{i,t}}$ using one-hot encoding corresponding to the first two digits of SIC code. Following the original paper, we construct 8 macroeconomic predictors denoted as x_{t}. The final covariate $\mathrm{z}_{\mathrm{i,t}}$ in the original work is calculated by

$$\mathrm{z}_{\mathrm{i,t}} = [[\mathrm{x}_{\mathrm{t}}, 1] \otimes \mathrm{c}_{\mathrm{i,t}}, \mathrm{s}_{\mathrm{i,t}}], \tag{3}$$

where $\otimes$ denotes the Kronecker product and P-dimensional vector is viewed as a $1 \times \mathrm{P}$ matrix. However, limited by the computational resources, we calculated the predictor variable $\mathrm{z}_{\mathrm{i,t}}$ by

$$\mathrm{z}_{\mathrm{i,t}} = [\mathrm{x}_{\mathrm{t}}, \mathrm{c}_{\mathrm{i,t}}, \mathrm{s}_{\mathrm{i,t}}]. \tag{4}$$

in our experiment while the code for (3) is also provided.

Data Splitting

- 18 years of training data(1957-1974), 12 years of validation data(1975-1986) and 34 years of test data(1987-2020).
- Refit models every year while increasing the training data by one year and maintain the same size of validation by rolling it forward until including all the data.
- After each refit, we use the model to predict the next year's excess returns.

Outline

Models

- OLS, OLS-3, ENet, PCR, PLS, GBRT, NN1, NN2, NN3, NN4 and NN5.
- Below we just discuss some details when conducting OLS-3 and NN models

OLS-3: Linear Regression with 3 Factors

- OLS-3 is the classical empirical asset pricing linear regression model using 3 factors as predictors.
- The three factors are the book-to-market(bm), the size(mvel1), and the monentum(mon1m, mon6m, mon12m, mon36m) while for monetum there are 4 characteristics calculated by data from the past 1, 6, 12, and 36 months, respectively.
- Thus, there are totally 6 predictors in OLS-3.

NN: Neural Networks

	NN1	NN2	NN3	NN4	NN5
neurons	32	32, 16	32,16,8	32,16,8,4	32,16,8,4,2

Table: Number of neurons in each hidden layer of NN models

- ReLU, Adam optimizer and MSE loss function.

Outline

Out-of-Sample R^2

For testing sample $\mathcal{T}_3$, the out-of-sample R^2 is defined as

$$\mathrm{R}^2_{\mathrm{oos}} = 1 - \frac{\sum_{(\mathrm{i},\mathrm{t})\in\mathcal{T}_3} (\mathrm{r}_{\mathrm{i},\mathrm{t}+1} - \hat{\mathrm{r}}_{\mathrm{i},\mathrm{t}+1})^2}{\sum_{(\mathrm{i},\mathrm{t})\in\mathcal{T}_3} \mathrm{r}^2_{\mathrm{i},\mathrm{t}+1}}. \tag{5}$$

- the denominator here is the sum of squared repsonce variables, not the sum of the squared residuals.

Result

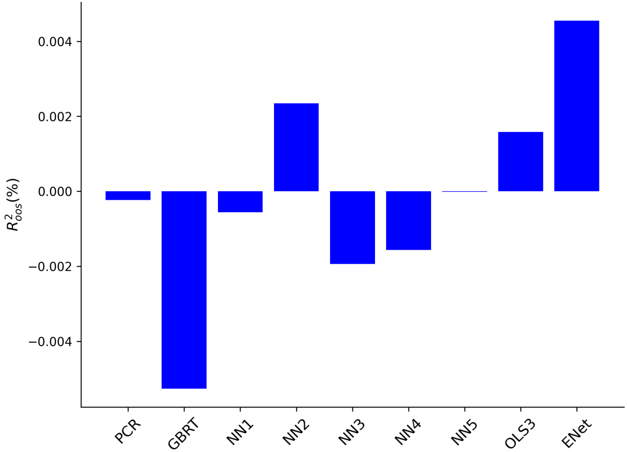
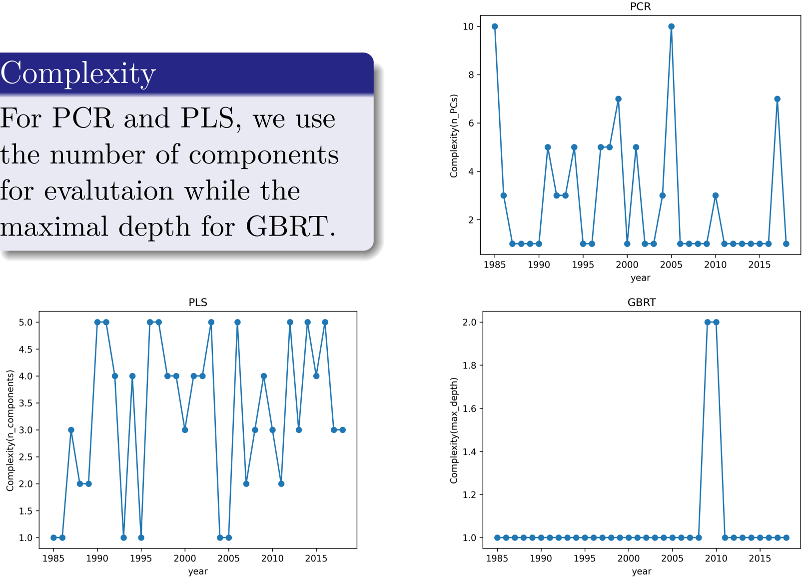


Figure: Out-of-sample R^2 performance of different models except OLS and PLS

Time-varing Model Complexity

Complexity

For PCR and PLS, we use the number of components for evalutaion while the maximal depth for GBRT.



Diebold and Mariano Test

Diebold and Mariano test is used to compare the out-of-sample predictive performance of model (1) versus model (2). Denote $\hat{e}_{i,t}^{(1)}$ and $\hat{e}_{i,t}^{(2)}$ as the prediction errors for stock i at testing month t of model (1) and model (2) respectively. Denote $\mathcal{T}_{3,t}$ as the set of stocks in the testing month t. Let

$$d_{12,t} = \frac{1}{|\mathcal{T}_{3,t}|} \sum_{i \in \mathcal{T}_{3,t}} \left((\hat{e}_{i,t}^{(1)})^2 - (\hat{e}_{i,t}^{(2)})^2 \right). \tag{6}$$

The test statistic is defined as $\mathrm{DM}_{12} = \bar{\mathrm{d}}_{12}/\hat{\sigma}_{12}$, where $\bar{\mathrm{d}}_{12}$ is the average of $\mathrm{d}_{12,\mathrm{t}}$ and $\hat{\sigma}_{12}$ is the Newey-West standard deviation, calculated by

$$\hat{\sigma}_{12} = \frac{1}{\mathrm{T}}\sqrt{\sum_{\mathrm{t}=1}^{\mathrm{T}}(\mathrm{d}_{12,\mathrm{t}} - \bar{\mathrm{d}}_{12})^2 + 2\sum_{\mathrm{l}=1}^{\mathrm{L}}\sum_{\mathrm{t}=\mathrm{l}+1}^{\mathrm{T}} \mathrm{w}_{\mathrm{l}}(\mathrm{d}_{12,\mathrm{t}} - \bar{\mathrm{d}}_{12})(\mathrm{d}_{12,\mathrm{t}-\mathrm{l}} - \bar{\mathrm{d}}_{12})} \tag{7}$$

where $\mathrm{w}_{\mathrm{l}} = 1 - \dfrac{\mathrm{l}}{\mathrm{L}+1}$ and $\mathrm{L} = \mathrm{T} - 1$.
And notice that $\hat{\sigma}_{12}$ is the autocorrelation-adjusted estimate of standard deviation of $\bar{\mathrm{d}}_{12}$.

	PCR	PLS	GBRT	NN1	NN2	NN3	NN4	NN5	OLS-3	ENet
OLS	**1.73**	**1.73**	**1.73**	**1.73**	**1.73**	**1.73**	**1.73**	**1.73**	**1.73**	**1.73**
PCR		**−2.52**	−1.56	−0.38	**3.28***	−0.72	−1.56	0.06	**3.09***	**6.35***
PLS			**2.27**	**2.54**	**2.59***	**2.66***	**2.44**	**2.39**	**2.64***	**2.86***
GBRT				**1.68**	**2.16**	1.20	1.30	1.40	**2.26**	**3.06***
NN1					**2.07**	−0.55	**−2.09**	0.29	**2.81***	**4.90***
NN2						−1.55	**−3.20***	**−2.31**	−0.52	**2.26**
NN3							−0.04	0.53	**1.70**	**3.22***
NN4								1.17	**4.38***	**6.28***
NN5									1.34	**3.88***
OLS3										**3.17***

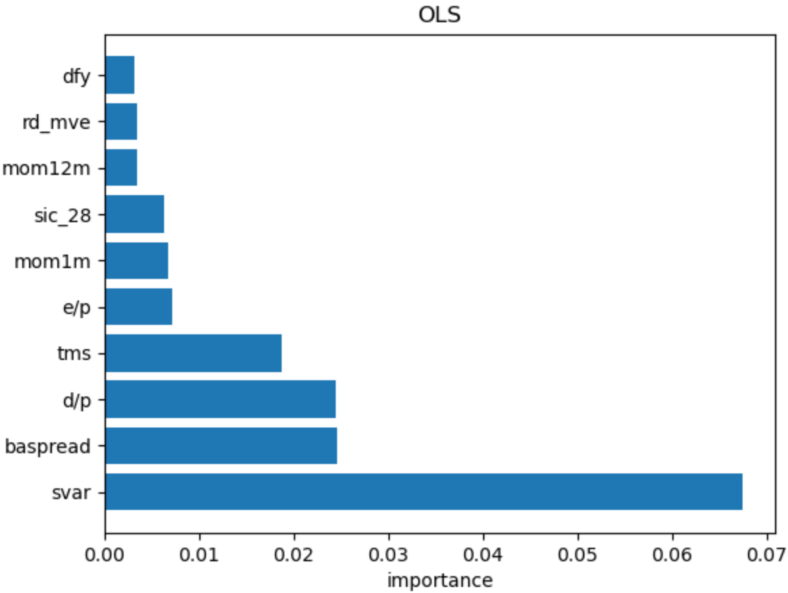
Figure: Diebold-Mariano test results

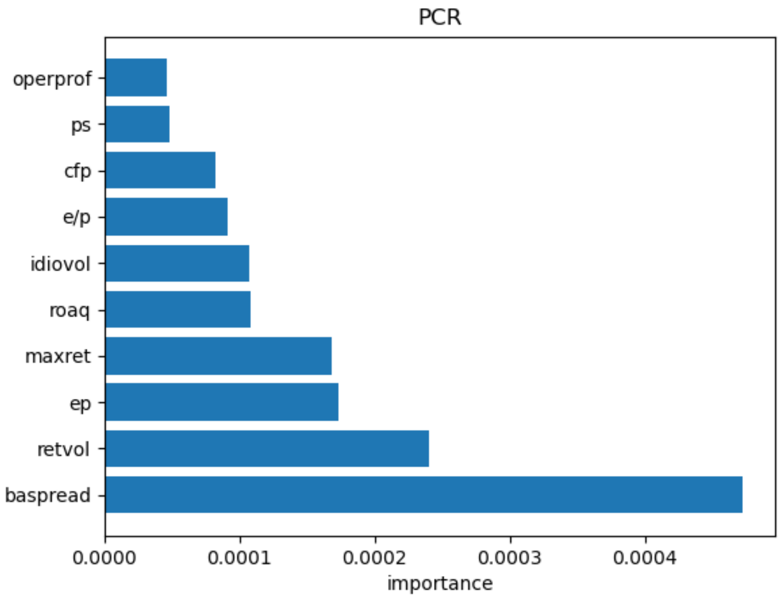
A positive value indicates the column model outperforms the row model while a negative value indicates the row model is better. The bold values indicate the difference is significant at 0.05 level. The star symbol indicates significance at 0.05 level for 10-way comparisons. The ENet model outperforms all other models significantly.

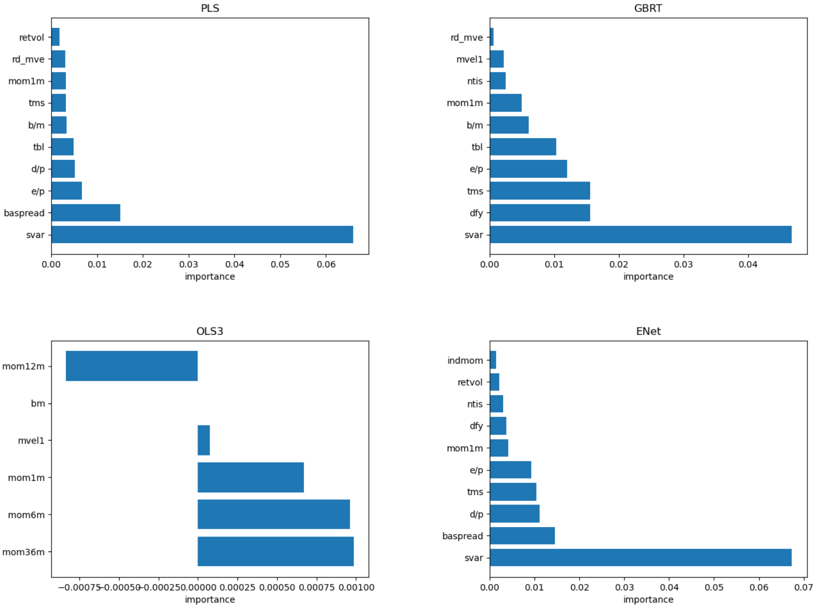
Outline

Variable Importance

We use permutation feature importance algorithm to study the variable importance. Generally, the bid-ask spread(baspread) and macroeconomic predictors(svar, d/p, etc) are important in our result.


OLS
dfy
rd_mve
mom12m
sic_28
mom1m
e/p
tms
d/p
baspread
svar
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
importance


PCR
operprof
ps
cfp
e/p
idiovol
roaq
maxret
ep
retvol
baspread
0.0000
0.0001
0.0002
0.0003
0.0004
importance


PLS
retvol
rd_mve
mom1m
tms
b/m
tbl
d/p
e/p
baspread
svar
0.00 0.01 0.02 0.03 0.04 0.05 0.06
importance
GBRT
rd_mve
mvel1
ntis
mom1m
b/m
tbl
e/p
tms
dfy
svar
0.00 0.01 0.02 0.03 0.04
importance
OLS3
mom12m
bm
mvel1
mom1m
mom6m
mom36m
−0.00075 −0.00050 −0.00025 0.00000 0.00025 0.00050 0.00075 0.00100
importance
ENet
indmom
retvol
ntis
dfy
mom1m
e/p
tms
d/p
baspread
svar
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07
importance

Outline

Conclusion

- replicated the main results of the original paper Empirical Asset Pricing via Machine Learning by Gu, Kelly, and Xiu (2018).
- confirm the findings of the original paper.
- provide more technical details.
- Limited by time and computational resources, we did not conduct the full replication of the original paper.
- More code in our GitHub repository[3].

[3] https://github.com/hdsfade/math5470-final-project

Contribution

- Zhen HOU is responsible for the code of PCR, performance evaluation, presentation and report writing.
- Jianda MAO is responsible for data preprocessing, training framework implementation, code of ENet, GBRT, NNs and variable importance.
- Xiaolong WANG is responsible for the code of OLS, OLS-3, PLS and running the experiments.