Paper Replication: Empirical Asset Pricing via Machine Learning

SHI, Runhao
Hong Kong University of Science and Technology
rshiaf@connect.ust.hk

1 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). In this report, 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 an elastic net penalty (ENet), principal components regression (PCR), partial least squares (PLS), random forests (RF), gradient boosted regression trees (GBRT), and neural networks (NN). For OLS, ENet, and GBRT, the Huber loss function is considered. Implementation details of these methods are discussed in Section 2. We evaluate the results of these 7 machine learning methods in Section 3, where a 'recursive performance evaluation scheme' containing 90-year time series data is used. The predictive trading characteristics include firm characteristics, sic code, and macroeconomic predictors. We evaluate the out-of-sample performance of each method and compare the importance of each characteristic contributing to the performance.

2 Methodology

In this section, we will give detailed information on implementing methods. We begin with the linear methods and then with the more complicated non-linear methods. Suppose we have P trading characteristics for each stock. The trading characteristics for stock i at time t are denoted as $\mathbf{z}^{(i,t)} \in \mathbb{R}^P$. We suppose the overall number of stocks is N and the time length is T. We denote the ith asset's excess return at time t as $r_{i,t+1}$.

2.1 Simple linear

The first method is the linear predictive regression model using ordinary least squares estimation (OLS) with Huber robust objective function. This method 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

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

2.2 Penalized linear

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

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

2.3 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 $\boldsymbol{\theta}_K \in \mathbb{R}^K$.

2.4 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).

2.5 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 diagram of the neural network is shown in Figure 1.

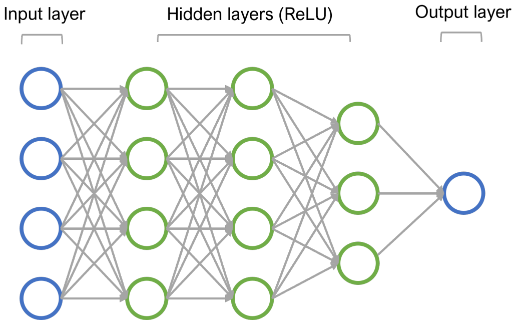

Figure 1: Neural networks architecture

3 Experiments

In this section, We run extensive experiments to evaluate the performances of different machine learning methods in real markets.

3.1 Performance evaluation

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

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

where $\mathcal{I}_3$ indicates the testing set.

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

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

$\hat{e}^{(i)}_{i,t+1}$ and $\hat{e}^{(2)}_{i,t+1}$ denote the prediction error for stock i at time t using method (1) and method (2), $n_{3,t+1}$ is the number of stocks at time $t+1$, $\bar{d}_{12}$ and $\hat{\delta}_{\bar{d}_{12}}$ denote the mean and Newey-West standard error of $d_{12,t}$ respectively.

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

3.2 Data preparation and model specification

The data we used are from Dacheng Xiu's website 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 characteristics and macroeconomics predictors are ranked period by period, which has been mapped into $[-1, 1]$ interval [Kelly et al., 2019, Freyberger et al., 2020]. Due to memory limitations, the total number of covariates we used in the linear model and regression trees is $94+8+74 = 176$. For the neural networks, the number of covariates we use is $94\times(8+1)+74 = 920$, which contains the stock characteristics, SIC code, and Kronecker product of stock characteristics and macroeconomics predictors. The hyper-parameters for all methods are shown in Table 1.

	OLS	PLS	PCR	ENet	RF	GBRT	NN
	+H			+H		+H	
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 1: Hyperparameters for all methods

We use a recursive performance evaluation scheme to evaluate the performance of different machine learning methods. The validating and testing are performing in a rolling-window basis, and the length of training data increases periodically. To be specific, the training dataset begins with 18 years sample (1957-1974), and the validating dataset takes the following 12 years sample (1975-1986). For each time, we use the training set and validating set to train our model, and test the performance using the following year data of the validating set. The length of training set increases one year each time while the length of validating remains the same size. We perform this process 30 times, and get 30 years predicting results.

3.3 Out-of-sample performance

We first show the out-of-sample R^2_{oos} of machine learning methods in Table 2 and Figure 2. We compare twelve models in total, including OLS with all covariates (we use 176 covariates for OLS in this report), OLS-3 (including preselect size, book-to-market, and momentum as the covariates), PLS, PCR, ENet, RF, GBRT, and NN (NN1,. . . , NN5). For OLS, ENet, and GBRT, we use the Huber robust loss function.

The first row of Table 2 shows R_{oos^2} for all stocks, the second row and the third row shows the R_{oos^2} of top-1000 stocks and bottom-1000 stocks by market equity each month respectively. For the second row and the third row, the estimated model uses all stocks but focuses on fits among the two subsamples.

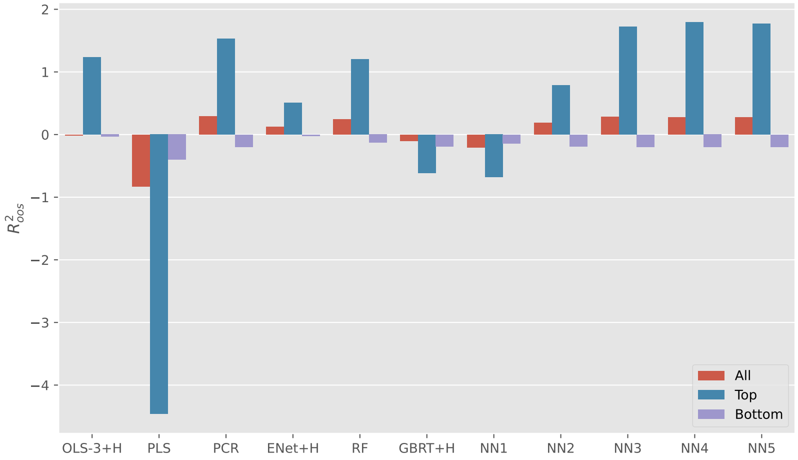

Figure 2: Monthly out-of-sample stock-level prediction performance (percentage R^2_{oos}).

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

OLS with Huber robust loss function performs the worst among all methods. The reason that accounts for the poor performance of OLS might be the lack of regularization which could be validated by the performance of OLS-3 and ENet. With fewer covariates or regularization, OLS-3 and ENet perform much better than OLS. ENet also achieves positive prediction accuracy in terms of All stocks and Top-1000 stocks. With the help of dimension reduction, PLS and PCR also raise the R^2_{oos} compared to OLS. As for PLS, it seems that R^2_{oos} is much lower than PCR. The reason might be the number of components in PLS is less than PCR, and one component may dominate the overall performance. Therefore, the weakness of choosing improper components has been amplified. As for the regression tree methods, RF has better performance with less computing time compared to GMRT. Neural networks are the best performing nonlinear method, also the best predictors in terms of all stocks and top-1000 stocks. From the results we can see that 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.

Table 3, 4 and 4 assess the statistical significance of differences among models. Bold numbers denote significance at the 5% level or better for each test. From the tables, we could see that 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 3: (All) Comparison of monthly out-of-sample prediction using Diebold-Marianon tests.

	OLS-3 +H	PLS	PCR	ENet +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
OLS+H	**10.25**	**8.5**	**10.49**	**9.44**	**10.47**	**9.32**	**10.36**	**10.47**	**10.53**	**10.59**	**10.55**
OLS-3+H	0	-7.26	**2.22**	-2.25	-0.04	-3.15	-2.59	-1.44	**3.16**	**3.48**	**3.21**
PLS	7.26	0	**7.19**	**6.78**	**6.86**	**5.48**	**3.74**	**6.62**	**7.21**	**7.32**	**7.22**
PCR	-2.22	-7.19	0	-2.5	-0.61	-3.29	-3.02	-2.25	**1.88**	**2.99**	**2.83**
ENet+H	**2.25**	-6.78	**2.5**	0	1.07	-2.06	-1.52	0.27	**2.98**	**3.09**	**2.99**
RF	0.04	-6.86	0.61	-1.07	0	-2.85	-2.13	-0.85	0.93	1.08	1.03
GBRT+H	**3.15**	-5.48	**3.29**	**2.06**	**2.85**	0	-0.36	1.64	**3.43**	**3.68**	**3.57**
NN1	**2.59**	-3.74	**3.02**	1.52	**2.13**	0.36	0	1.92	**3.18**	**3.32**	**3.22**
NN2	1.44	-6.62	**2.25**	-0.27	0.85	-1.64	-1.92	0	**2.81**	**2.99**	**2.93**
NN3	-3.16	-7.21	-1.88	-2.98	-0.93	-3.43	-3.18	-2.81	0	0.75	0.41
NN4	-3.48	-7.32	-2.99	-3.09	-1.08	-3.68	-3.32	-2.99	-0.75	0	-0.54
NN5	-3.21	-7.22	-2.83	-2.99	-1.03	-3.57	-3.22	-2.93	-0.41	0.54	0

Table 4: (Top) Comparison of monthly out-of-sample prediction using Diebold-Marianon tests.

	OLS-3 +H	PLS	PCR	ENet +H	RF	GBRT +H	NN1	NN2	NN3	NN4	NN5
OLS+H	**6.68**	**8.54**	**6.94**	**6.98**	**7.26**	**6.96**	**6.84**	**6.84**	**6.9**	**6.83**	**6.79**
OLS+H	**6.68**	**8.54**	**6.94**	**6.98**	**7.26**	**6.96**	**6.84**	**6.84**	**6.9**	**6.83**	**6.79**
OLS-3+H	0	-0.3	-0.03	0.75	0.29	-0.74	0.04	-0.01	-0.13	-0.17	-0.1
PLS	0.3	0	0.33	0.81	0.82	-0.21	0.39	0.33	0.21	0.16	0.22
PCR	0.03	-0.33	0	0.65	0.65	-0.44	0.33	0.11	-0.59	-1.06	-0.62
ENet+H	-0.75	-0.81	-0.65	0	-0.19	-1.3	-0.59	-0.68	-0.9	-0.97	-0.83
RF	-0.29	-0.82	-0.65	0.19	0	-0.89	-0.52	-0.58	-0.88	-0.99	-0.82
GBRT+H	0.74	0.21	0.44	1.3	0.89	0	0.51	0.46	0.35	0.32	0.36
NN1	-0.04	-0.39	-0.33	0.59	0.52	-0.51	0	-0.22	-0.99	-1.29	-1.01
NN2	0.01	-0.33	-0.11	0.68	0.58	-0.46	0.22	0	-0.73	-0.92	-0.66
NN3	0.13	-0.21	0.59	0.9	0.88	-0.35	0.99	0.73	0	-0.36	0.14
NN4	0.17	-0.16	1.06	0.97	0.99	-0.32	1.29	0.92	0.36	0	0.76
NN5	0.1	-0.22	0.62	0.83	0.82	-0.36	1.01	0.66	-0.14	-0.76	0

Table 5: (Bottom) Comparison of monthly out-of-sample prediction using Diebold-Marianon tests.

3.4 Comparison of variable importance

We now investigate the relative importance of individual covariates for the performance of each model using the variable importance measures. Figure 3 demonstrate the resultant importance of the top-20 characteristics and macroeconomic predictors for each model. Variable importance within the model is normalized to sum to one. Figure 3 shows that variable importance magnitudes for PLS, ENet and RF are highly skewed toward corporate investment (cinvest).

4 shows the heatmap of overall rankings of characteristics for all models. The characteristics are ordered with the highest total ranks on top and the lowest at the bottom. From the heatmap, we could see that the high-ranking characteristics are cinvest, rd, mom12m, stdcf, stdacc, and bm.

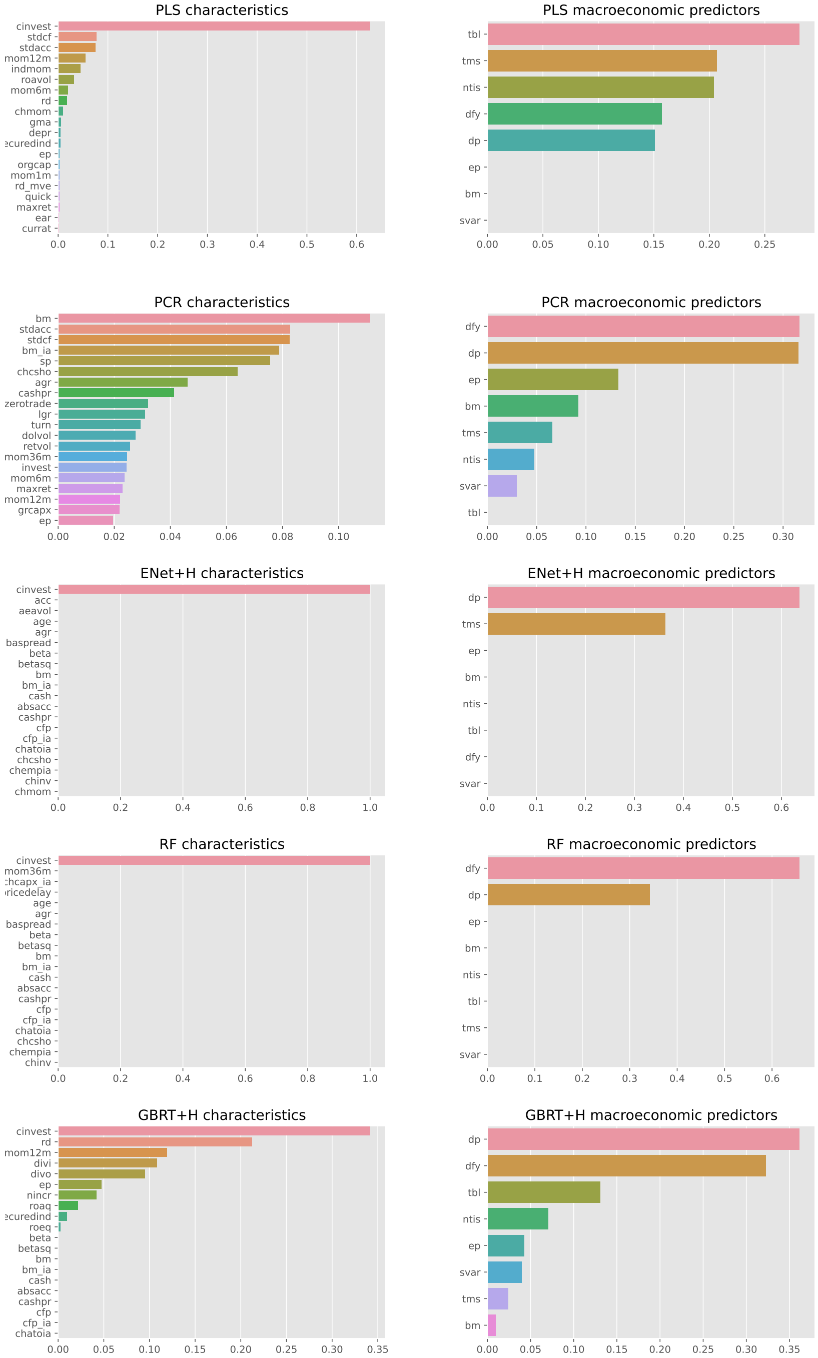

Figure 3: Variable importance by model

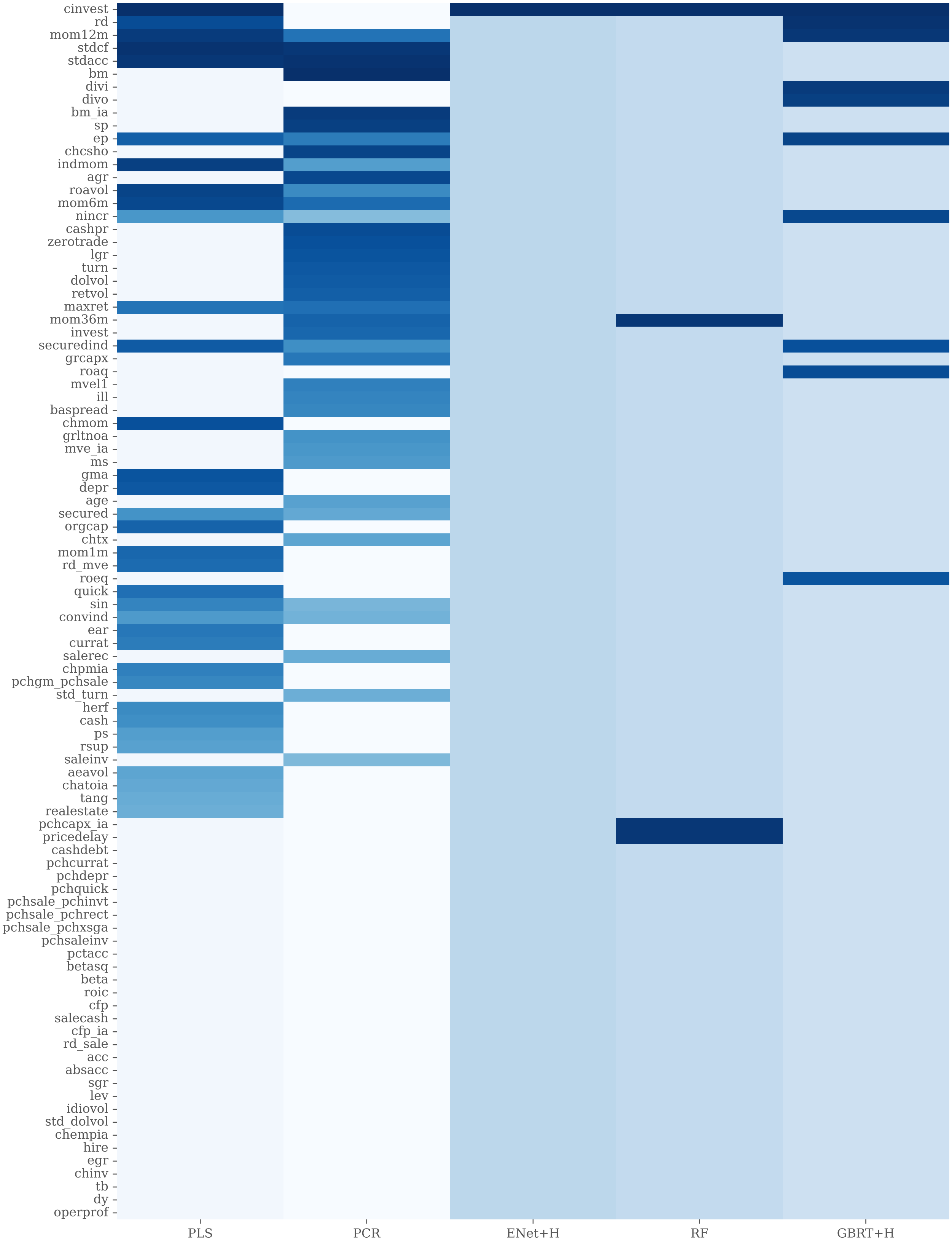

Figure 4: Characteristic importance.

Figure 5 and Table 6 shows the R^2_{oos}-based importance measure for each macroeconomic predictor, where variable importance within each model is normalized to sum to one. From Figure 5, we could see that the dividend-price ratio (dp) and default spread (dfy) are critical predictors. From Table 6, we could see that stock variance (svar) is the least informative macroeconomic predictor compared to others. The reason account for this might be stock variance is noisier than other variables.

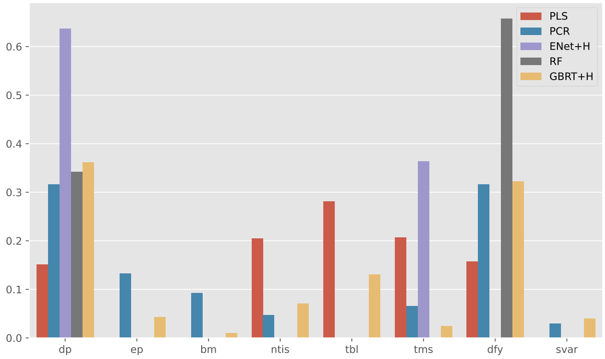

Figure 5: Variable importance for macroeconomic predictors.

	PLS	PCR	ENet+H	RF	GBRT+H
dp	0.15	0.32	0.64	0.34	0.36
ep	0	0.13	0	0	0.04
bm	0	0.09	0	0	0.01
ntis	0.2	0.05	0	0	0.07
tbl	0.28	0	0	0	0.13
tms	0.21	0.07	0.36	0	0.02
dfy	0.16	0.32	0	0.66	0.32
svar	0	0.03	0	0	0.04

Table 6: Variable importance for macroeconomic predictors.

4 Conclusion

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

References

Joachim Freyberger, Andreas Neuhierl, and Michael Weber. Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377, 2020.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

Bryan T Kelly, Seth Pruitt, and Yinan Su. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524, 2019.