Introduction to “Empirical Asset Pricing via Machine Learning”

By

Shihao Gu
University of Chicago

Bryan Kelly
Yale University, AQR Capital Management, and NBER

Dacheng Xiu
University of Chicago Booth School of Business

Background

- Risk premium is difficult to measure: market efficiency forces return variation to be dominated by unforecastable news that obscures risk premiums.
- Machine learning accommodates a far more expansive list of potential predictor variables, which enables gains that can be achieved in prediction and identifies the most informative predictor variables.

This paper uses machine learning methods to **predict asset's excess return->regression problem**

- Linear models: OLS, elastic net
- Dimension reduction: PLS, PCR
- Generalized linear model
- Tree models: Gradient boosted regression tree, random forest
- Neural networks

Experiment preparation

Data and feature

Monthly total individual equity returns for all firms listed in NYSE, AMEX, NASDAQ.
~30,000 stocks over 60 years from March 1957 to December 2016.

Characteristics including:

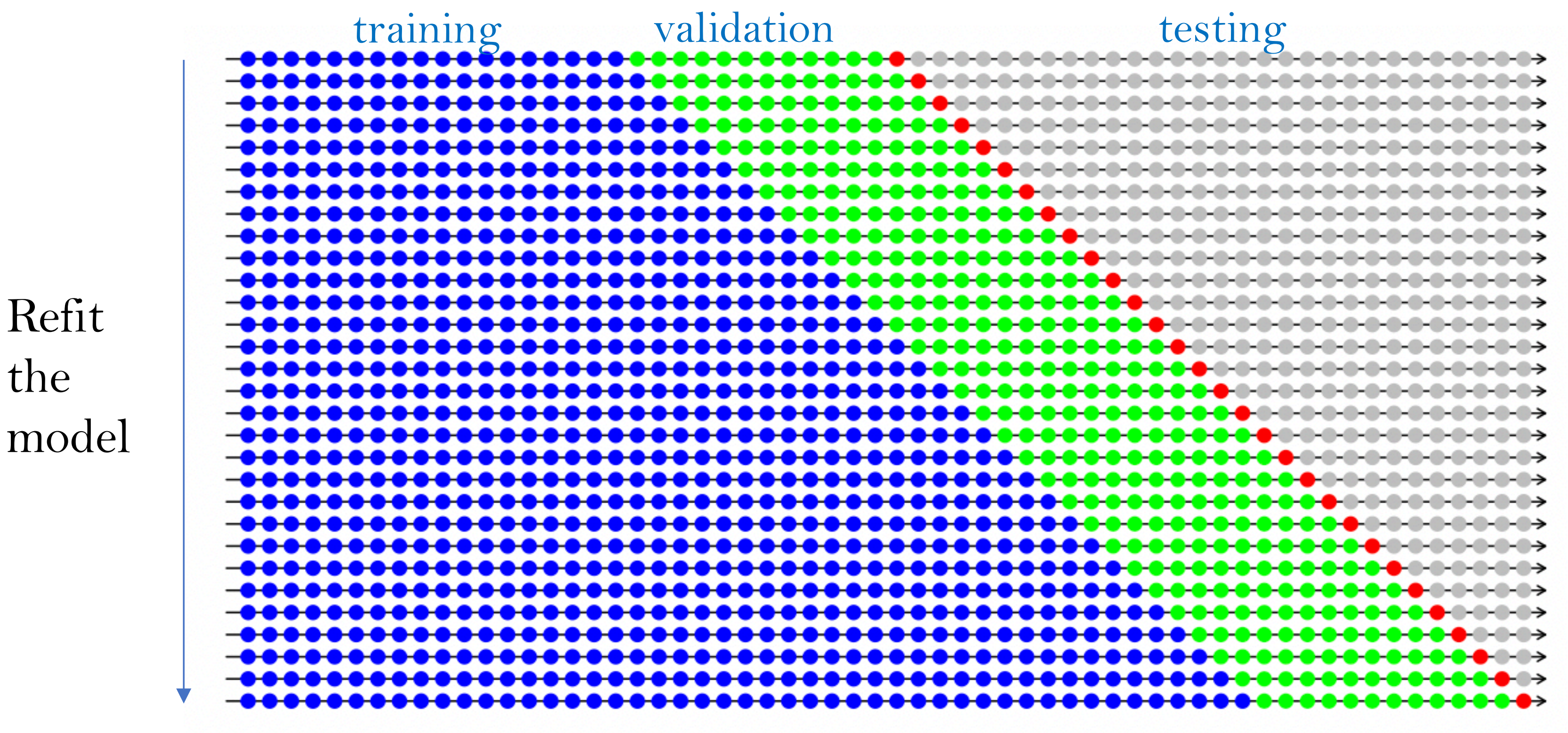
- 94 firm characteristics
- 8 macroeconomic predictors
- 74 industry dummies

More details are described in Sec. 2.1

Experiment preparation

Divide the 60 years of data into 18 years of training sample (1957-1974), 12 years of validation sample (1975-1986), and the remaining 30 years for out-of-sample testing (1987-2016).

Adopt a **recursive performance evaluation scheme**.



training

validation

testing

Refit the model

More details are described in Sec. 2.1

Objective function =>Tune the model's parameter on the training set

- *Mean Squared Error (MSE) loss*

Basic formula:

Label,
i.e., real return

Model's
prediction

$$\mathcal{L}(\theta)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\boxed{r_{i,t+1}}-\boxed{g(z_{i,t};\theta)}\right)^2$$

$i = 1, \ldots, N$: stock index
$t = 1, \ldots, T$: month index

Evaluation function =>Evaluate the models' performance on the testing set

- *Out-of-sample R^2*

Basic formula:

Label,
i.e., real return

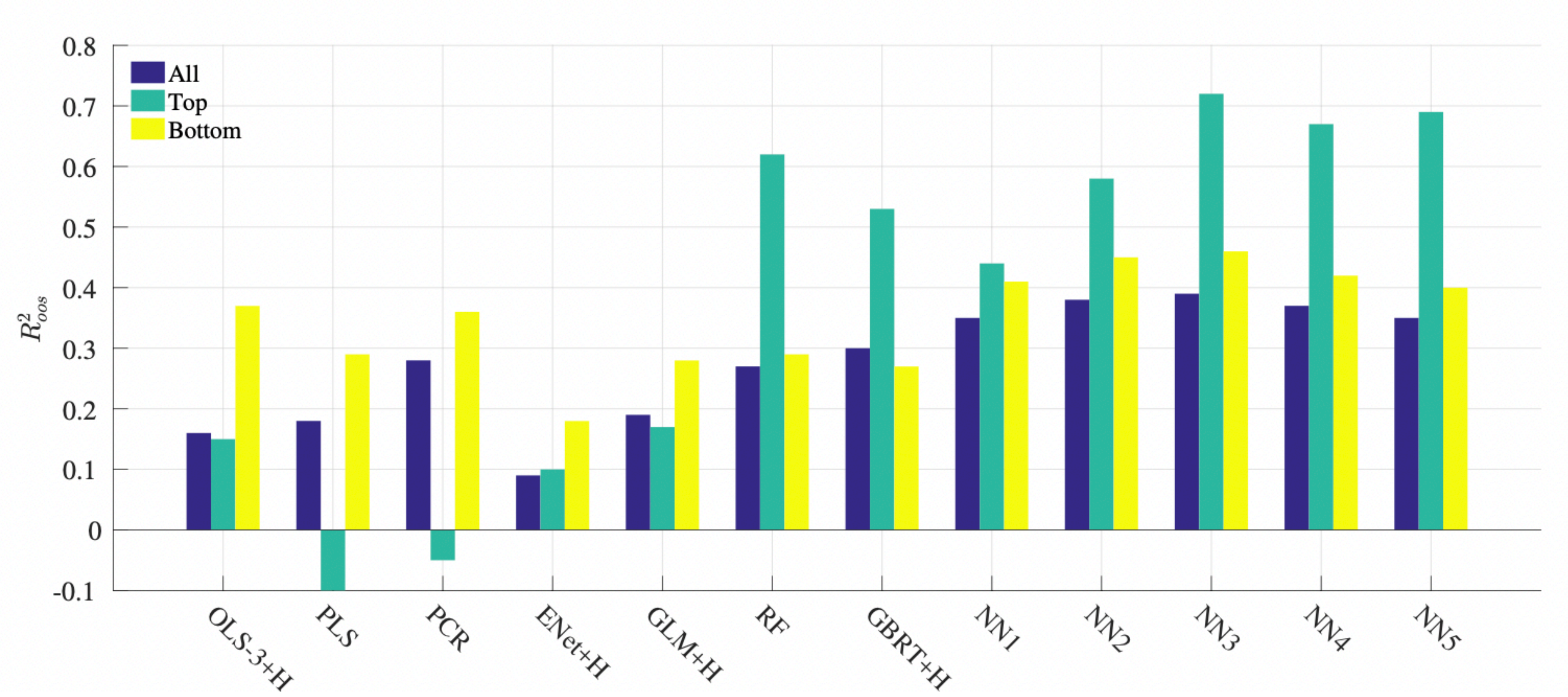
Model's
prediction

$$R^2_{\text{oos}}=1-\frac{\sum_{(i,t)\in\mathcal{T}_3}\left(\boxed{r_{i,t+1}}-\boxed{\widehat{r}_{i,t+1}}\right)^2}{\sum_{(i,t)\in\boxed{\mathcal{T}_3}} r^2_{i,t+1}}$$

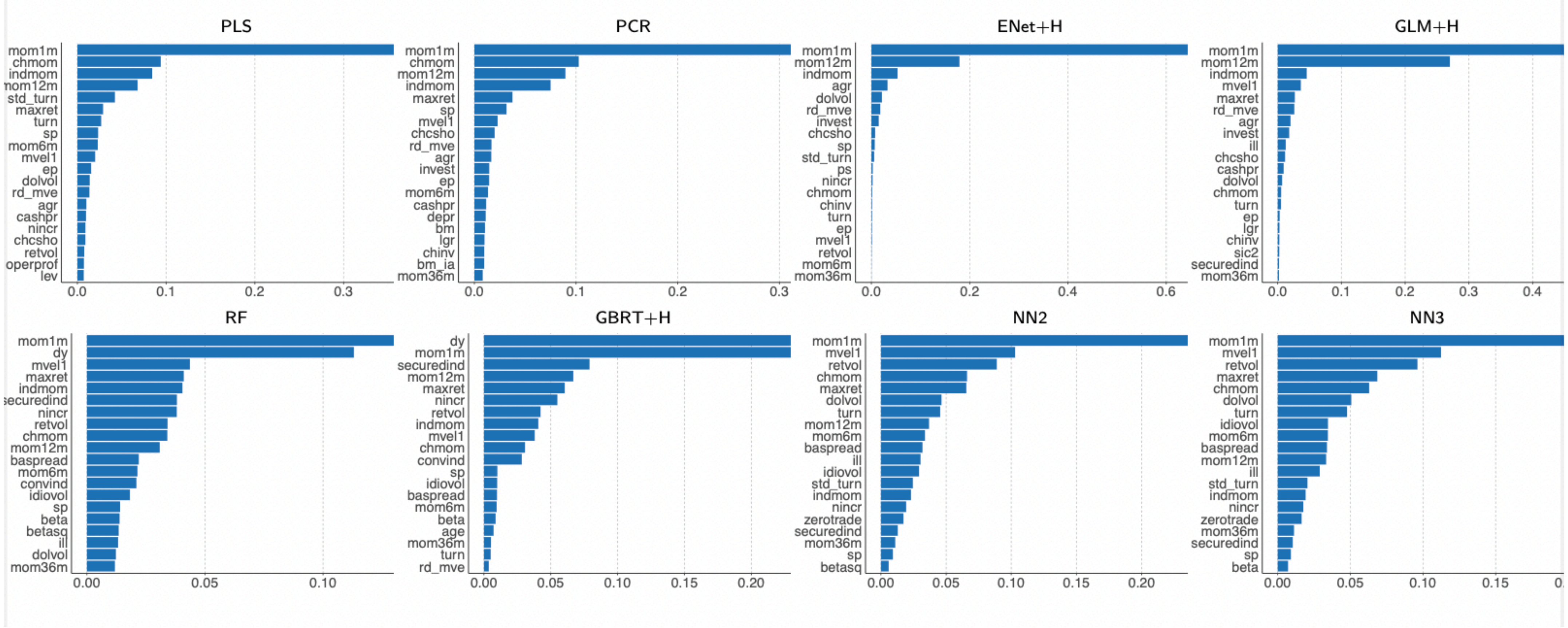
Testing set

Some results:

- **Individual Stock Returns Prediction**



Characteristic Importance



Requirements for replication

- Data Preparation (Adopt the **recursive performance evaluation scheme)**

- Model selection
  - ➢ Replicate **at least 6 models** (Hints of parameter chosen are presented in the paper).

- Results analysis
  - ➢ Variable importance
  - ➢ Model performance comparison and analysis

□ You **do not need** to replicate the results of **section 2.4**: Portfolio forecast.

□ Supplementary material can be helpful to you.