MAFS6010Z Project 2: Paper replication I: Empirical Asset Pricing via Machine Learning

LI Aoran, MA Yijia, WENG Langting, ZHOU Tianying

Financial Mathematics - HKUST

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Abstract

This report is an replication of the paper "Empirical Asset Pricing via Machine Learning". We reproduce 6 machine learning methods, including linear regression(**OLS**, **OLS-3**, **Elastic Net**), dimension reduction(**PLS**, **PCR**), **Random Forest** and **GBRT**.As for the evaluation part, we present the comparison of machine learning techniques in terms of their $R^2_o os$ and DM test outcomes, demonstrate the models' complexity, and analyze the variable importance for models. Finally we identify the best-performing methods and obtain the conclusion.

Data Processing

We select 60 years of data from 01/01/1961 to 31/12/2020, generate 8 macroeconomic predictors specified in the paper, and then deal with missing values.

- 94 stock-level predictive characteristics
- 74 industry dummies corresponding to SIC code
- 8 macroeconomic predictors

Comparison of Machine Learning Techniques

We use **recursive performance evaluation scheme** to

	OLS-3+H	PLS	PCR	ENet	RF	GBRT
All	-0.016	0.069	-0.131	0.392	0.346	0.517
Top1000	-0.0674	0.139	0.0971	0.169	0.448	0.516
Bottom1000	-0.013	-0.0619	0.115	0.312	0.268	0.338

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

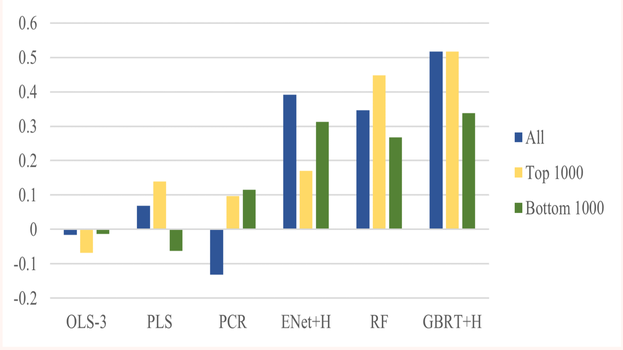

Figure 1. Annual out-of-sample stock-level prediction performance

- We conduct our analysis at the annual horizon.
- It is obvious that linear models does not have a good performance via $R^2_o os$ value. There is an increasing tendency, which means machine learning models are better in predictive performance.

Time-varying model complexity

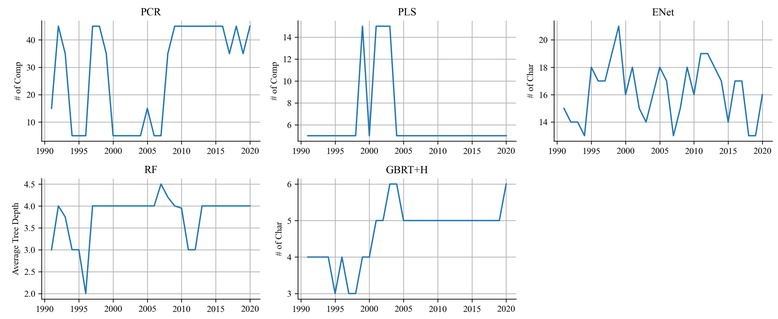

Figure 2. Time-varying model complexity

This figure demonstrates the models' complexity for PCR ,PLS ,elastic net(ENet),random forest (RF), and gradient boosted regression trees (GBRT) in each training sample of our 30-year recursive out-of-sample analysis. For ENet, we report the number of features selected to have non zero coefficients; for PCR and PLS, we report the number of selected components; for RF, we report the average tree depth; and, for GBRT+H, we report the number of distinct characteristics entering into the trees. It summarizes the complexity of each model at each reestimation date. The upper left panel shows the number of features to which elastic net assigns a nonzero loading. It shows that PCR typically uses 10 to 40 components in its forecasts.PLS, finds less reliable components of the early sample and late sample, but eventually settles on six to fourteen components. Random forests generally estimate shallow trees, with two to five layers on average. To quantify the complexity of GBRT, we report the number of features used in the boosted tree ensemble at each reestimation point.

CharacteristicsImportance

We investigate the characteristic importance for performance of each model using the importance measures. We rank stock-level characteristics in terms of overall models.Characteristics are ordered based on the sum of their ranks overall models,with the most influential characteristics on the top and the least influential on the bottom. Columns correspond to the individual models,and the color gradients within each column indicate the most influential(dark blue) to the least influential(white).

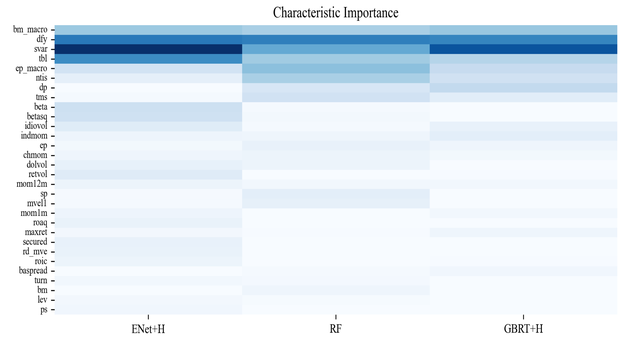

Figure 3. Characteristics Importance

Macroeconomic Predictors

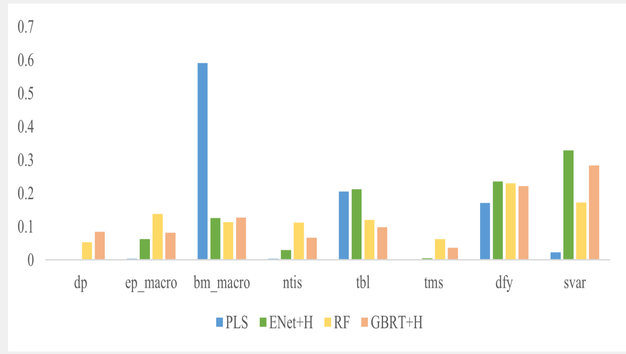

Figure 4. Macroeconomic predictors analysis

Future Work

Future Work

- In OLS+H model ,we use stock-level covariates in subsamples that include only the top_1000 stocks or bottom_1000 stocks, since the complete data is too large to deal with due to limited computational power,which is what we want to improve in our future research.
- Ideal settings of parameters for RF and GBRT models like n_estimators, max_depth can't be reached in this replication because of the computer limited computational power,so the R^2 and complexity results are different from the paper, so we want to try some more aggressive hyperparameter setting in validation process in future work.
- In this replication, the best models are tree-based models. We can conduct some other tree-based models like LightGBM and XGBoost in future work.
- Original paper aims to use machine learning methods to analyze predictability of individual stock returns and compare portfolio forecasting performance, which is what we want to do in our future research.

References

[1] Gu, Shihao and Kelly, Bryan T. and Xiu, Dacheng, Empirical Asset Pricing via Machine Learning (September 13, 2019).
[2] A Comprehensive Look at the Empirical Performance of Equity Premium Prediction (with Ivo Welch), July 2008, Review of Financial Studies 21(4) 1455–1508.

Contribution

Model Construction: ZHOU Tianying、MA Yijia, WENG Langting
Report Writing：LI Aoran, MA Yijia, WENG Langting, ZHOU Tianying