Project 2: Paper Replication Study
Empirical Asset Pricing via Machine Learning

Sixian HUO shuo@connect.ust.hk

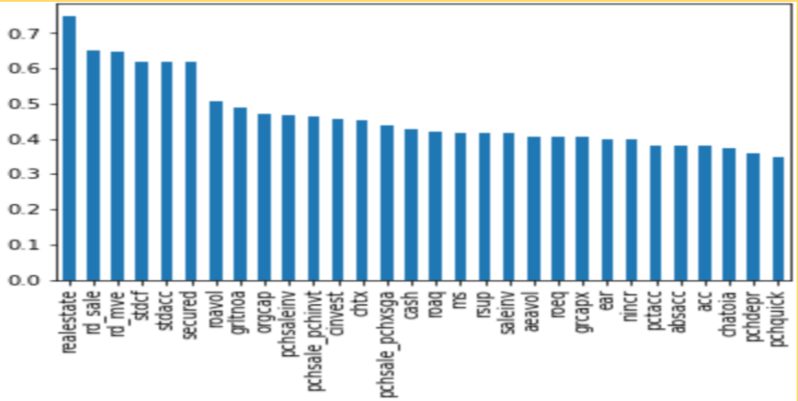
1. Introduction

This paper shows a comparative analysis of machine learning methods for the canonical problem of empirical asset pricing. They compare thirteen models in total and find that the nonlinear model outperforms the linear model. In this project, I try to replicate 6 methods they used, including OLS, ENET, PLS, PCR, RF and GBRT. Then I also find that GBRT and RF had a better performance with the size and momentum features.

2. Data Preprocessing

DATE RANGE
Focus on data between 1957 and 2016.

MISSING VALUE
There are many missing value in the dataset. Using forward and backward methods to fill the missing value and drop the remaining data with missing value.



Percentage of Missing Value

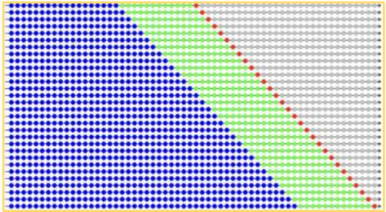
3. Performance Evaluation

Different from the traditional R2 score, this paper uses a method without demeaning. It gives a higher benchmark to individual stock returns.

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

4. Recursive Evaluation Method

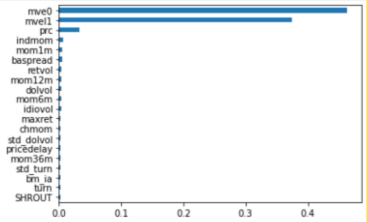
To avoid refitting models, they roll the training sample forward to include the most 12 months. This can be done by a generator in python, which can reduce the memory consumption.



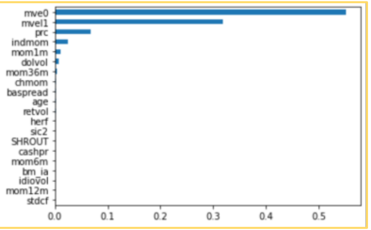
Recursive Evaluation Method

5. Replication

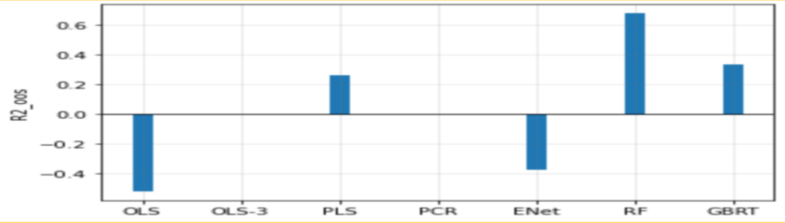
Using recursive evaluation method to generate dataset, then put them into different models and calculate their R2_oos score. We can get the performance and variable importance.



RF



GBRT



Score of Different models

6. Conclusion

We can find that the nonlinear models, RF and GBRT had a better score than other linear models. It is the same as the paper. And in terms of variable importance, mve and mom have a higher ranking, which represents that the liquidity is a powerful feature to predict returns.

The replication result is not completely consistent with the paper. Because I didn't add the 8 macroeconomic predictors due to the limitation of memory and I'm a little bit confused about them.

7. References

Shihao Gu, Bryan Kelly and Dacheng Xiu "Empirical Asset Pricing via Machine Learning" (2020)

8. Contribution

DATA CLEANNING, REPLICATE, REPORT
Sixian HUO, 20810798