MATH 5470- Project: Empirical Asset Pricing via Machine Learning

XIONG Wei[1], LIU Chen[1] , WANG Zhe[2] and JI Wen[2] {wxiongae, cliudh, zwangec, wjiac}@connect.ust.hk

[1]: Department of Mathematics, HKUST [2]: Department of Computer Science and Engineering, HKUST

https://www.bilibili.com/video/BV1nL4y1c71g?spm_id_from=333.999.0.0

1. Introduction

Empirical Asset Price is an essential problem in finance, which is helpful to measure equity risk premiums. In this project, we implement the work of "Empirical Asset Price via Machine Learning". After data cleaning, a set of models, e.g., OLS, GBDT, LGB NN, etc., are constructed to obtain the predictive results. We calculate R^2_{oos} to measure the performance of each model. Furthermore, we report the resultant importance of the top-20 stock-level characteristics.

2. Data Processing

Predictors Construction: As the paper suggested, we first construct several macroeconomic predictors that are not provided in the original dataset.

Missing Value Elimination: we handle the missing values by two strategies: 1 If the percentage of missing value < 50%, we replace the it with the average value. 2 If the percentage of missing value > 50%, we replace the missing value with the mode value.

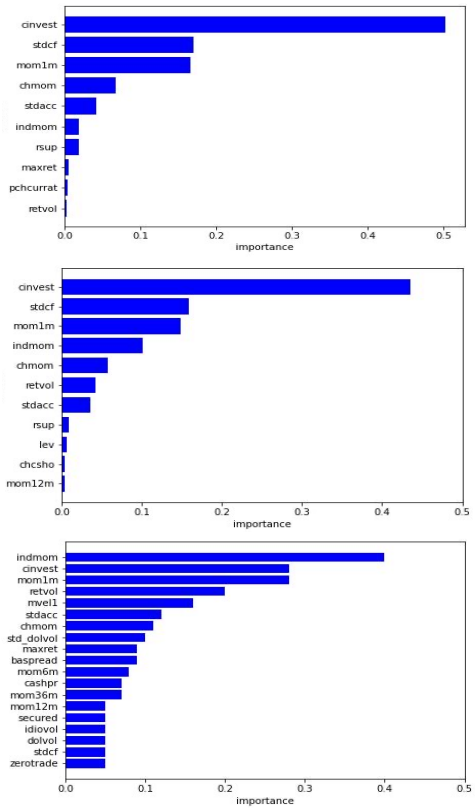
Data Normalization: We also perform standard normalization procedure with sample mean and deviation to facilitate training.

3. Model Replication and Performance

We implement all the benchmark models reported in the original paper, including OLS, OLS3, Ridge, Lasso, Huber, ENET, PLS, PCS, RF, GBDT, LGBM, and a series of NN under different settings. And R^2_{oos} is calculated as the performance metric. According to the results reported in **Table 1**, we can conclude that: 1 nn4 can obtain the best performance, we think it is benefited by the non-linear structure 2 Compared with the methods based on Linear Regression and tree, the per--formances Of the Non-linear method are more promotive.

$$R^2_{oos} = 1 - \frac{\sum_{(i,t)\in\tau_3}\left(r_{i,t+1} - \hat{r}_{i,t+1}\right)^2}{\sum_{(i,t)\in\tau_3} r^2_{i,t+1}}$$

4. Characteristic Importance



From top to bottom are the variable importance visualization of RF, GBDT and LGBM, respectively. We analyzed the characteristic importance of the tree-based methods, we can obtain similar conclusions to the original paper. The tree-based methods are more democratic, drawing predictive information from a broader set of characteristics.

Table 1. Performance (%)

OLS	-7.70	**RF**	-0.52
OLS3	-0.49	**GBDT**	-0.16
Ridge	-7.10	**LGB**	0.08
Lasso	0.21	**NN1**	0.27
Huber	-128.08	**NN2**	0.28
ENET	0.21	**NN3**	0.28
PLS	-1.09	**NN4**	0.36
PCS	0.20	**NN5**	0.24

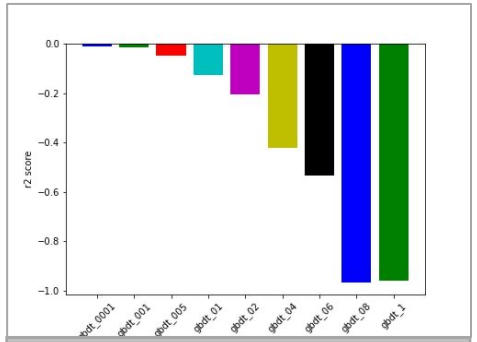


Fig1. Performance of GBDT with different lr

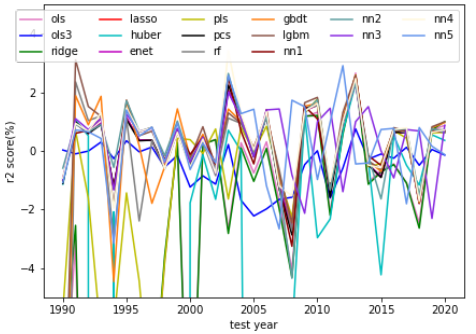


Fig2. Performance with different test year

5. Analysis

The characteristic importance: The figures in session 4 demonstrate that models are generally in close agreement regarding the distribution of the most influential stock-level predictors. Especially, the Recent Price Trends have the greatest impact (e.g., mom1m, mom12m, chmom, indmom).

Performance of GBDT with different learning rate: The performance of the GBDT model is very sensitive to changes in the learning rate, when the learning rate is 0.001, the GBDT model achieves the best performance.

Performance with different test year: As shown in **Fig 2**, over our 30- year out-of-sample period, compare with tree-based and NN-based methods, the linear models (e.g., OLS, Ridge, PLS) exhibit high volatility.

6. Conclusion and Further Improvement

In this project, we do the paper replication for the problem of empirical asset pricing via machine learning. In summary, for the R^2_{oos} performance, nn4 performs best. For the characteristic importance, the tree-based methods are democratic.

There is still a lot of room for improvement: we observed that the test loss is general worse than the training loss, possibly due to the dynamic nature of the financial data, where training distribution and the test distribution can different. In this case, the techniques from out-of-domain research can be beneficial.

7. References

Gu S et al. Empirical asset pricing via machine learning[J]. The Review of Financial Studies, 2020, 33(5): 2223-2273.

8. Contribution

Data processing: LIU Chen (ID:20809024)

Model Replication:

WANG Zhe (ID:20550960), XIONG Wei (ID:20807868)

Poster Report and Results Analysis: JI Wen (ID:20842064)