MAFS6010Z (L1) - Artificial Intelligence in Fintech Project 2--Paper Replication I: Empirical Asset Pricing via Machine Learning

HUANG, Yuxin; LEI, Yunxin; AN, Tianyuan; LIN, Fengshan; LIU, Zongxuan; : Department of Mathematics, HKUST

1. Introduction

We replicated eight of the models used in the article: OLS, Elastic Net, PLS, PCR, Generalized Linear Models, Gradient Boosting Trees, Random Forest, and Neural Networks. Inspired by the last project, we tried two gradient boosting tree models: LightGBM and XGB.

2. Methodology - Hyperparameter tuning

Simple Splitting

Due to computational power constraints, our hyperparameter tuning is based on the data of the 1000 stocks with the highest feature mvel1, with a fixed time window of 18 years training set, 12 years validation set, and 30 years test set.

Performance Evaluation

The selection of the best parameters is based on the comparison of out-of-sample R-squared:

$$R^2_{OOS} = 1 - \frac{\sum_{(i,t)}(r_{i,t+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)} {r_{i,t+1}}^2}$$

3. Model Fitting

- Firstly, we used Parameter Grid to perform a full parallelized parameter search and generate a list of parameters.
- Secondly, for each set of parameters, we fitted the model with training data, made predictions on validation data, and calculated out-of-sample R-squared.
- Finally, by comparing the out-of-sample R-squared, we got the best model and set of parameters.

3' Best Parameter of Models

- OLS & OLS3: Huber objective function;
- PLS: Components (1);
- PCR: Principal components(5), MSE objective function;
- GLM: nodes(3), coefficient of GroupLasso regularization(0.0001), and coefficient of Lasso regularization(0.0001);
- Random Forest: estimator(300), max depth(6), max features(100);
- LightGBM: estimator(100),max depth(2), learning rate(0.1);
- XGB: estimator(500), max depth(2), learning rate(0.01) ;
- NN: learning rate(0.01)

4. Performance Evaluation(PE)

- Apply the best set of parameters on the model to make predictions on testing data; and compare out-of-sample R-squared of different models.
- In particular, there are several implementations of Elastic Net. The data in the table below is the one with the best performance.

	In-sample R-squared	Out-of-sample R-squared	In-sample MSE	Out-of-sample MSE
OLS	3.11%	-50.96%	0.0096	0.0081
OLS3	0.79%	-68.77%	0.0098	0.0090
EN	0.00%	1.30%	0.005	0.006
PLS	6.92%	-41.61%	0.0092	0.0076
PCR	0.030%	-19.26%	0.0099	0.0064
GLM	1.61%	0.72%	0.08	0.07
lgbm	19.41%	1.26%	0.009	0.009
RF	19%	-0.07%	0.006	0.011
XGB	14.42%	0.14%	0.007	0.011
NN1	-15.29%	-337.46%	0.007	0.041
NN2	-42.41%	-24.73%	0.009	.012
NN3	-5.16%	-13.75%	0.007	0.011
NN4	-65.5%	0.54%	0.010	0.009
NN5	-19.09%	0.40%	0.007	0.009

TABLE1: In-sample & out-of-sample R-squared & MSE

5. Feature Importance(FI)

- For tree models (LightGBM, XGB, and Random Forest), feature importance can be obtained directly from the regressor. However, with respect to other models, it needs other methods to obtain the feature importance.
- Take Neural Networks as an example. In reference to the paper and some other resources, we calculate the reduction in panel predictive R-squared from shuffling values of predictor j, while holding the remaining model estimates fixed.
- Tables of feature importance can be found in the notebook attached.

6. Expanding Windows on PE and FI

- Take NN1 as an example. We first trained the NN1 model with all features on 30 stocks and calculated the feature importance. We then sorted the features descendingly and selected the top 100 features to train the Neural Network models on 500 stocks. For each model, we first trained it on the whole sample(48 years of training set and 12 years of validation set) and got the best parameters. Then we tuned the model by adopting the recursive performance evaluation scheme with expanding windows and fixed parameters.
- More details can be found in the notebook attached.

7. Conclusion

- **Model Comparison**: we find that tree models the best under the recursive performance evaluation scheme. They have the highest R-squared, both in-sample and out-of-simple.
- **Feature importance**: we find some variables weigh a lot among all models. **Percent accruals** (pctacc) implies the company's debt position. **12-month momentum** (mom12m) reflects the long-term trend of the stocks' price. **Industry-adjusted book-to-market** (bm_ia) illustrates the market's assessment of the company.

8. References

- Notebooks and codes on Website, especially Yanlin Bao, PhD in Business(Finance) student at Singapore Management University;
- Shihao Gu, Bryan Kelly, Dacheng Xiu, Empirical Asset Pricing via Machine Learning, *The Review of Financial Studies*, Volume 33, Issue 5, May 2020, Pages 2223–2273

9. Contribution

- **Model Fitting & Performance Evaluation & Expanding windows:**
  - **Lin: OLS, OLS3, PLS, PCR;**
  - **Liu: PLS, PCR, Elastic Net;**
  - **Huang: GLM;**
  - **An: LightGBM, XGB, Random Forest;**
  - **Lei: NN1-NN5;**
- **Code Direction: Lei;**
- **Report Integration: An.**