Paper Replication: Empirical Asset Pricing via Machine Learning

Yuxuan Chen yx.chen@connect.ust.hk

Department of Mathematics, HKUST

Introduction

Empirical Asset Price is an essential problem in finance, which is helpful to measure equity risk premiums. In this project, we replicate some essential work in the paper 'Empirical Asset Pricing via Machine Learning'. In specific, we replicate some of the methods, evaluate their performance using a 'recursive performance evaluation scheme', and further analyse the variable importance.

Data Processing

- **Macroeconomic Predictors and Excess Returns** Macroeconomic predictors and excess returns are not provided in the original dataset and we need to construct them by ourselves. Following [1], we construct 8 macroeconomic predictors according to the description in [2]. For excess returns, we download CRSP returns from WRDS and calculate excess returns by subtracting Treasury-bill rates from CRSP returns.
- **Missing Characteristics Processing** Following [1], we replace missing values with the cross-sectional median at each month for each stock, respectively.
- **Data Normalization** We perform standard normalization procedure with sample mean and deviation to facilitate training.

Model Replication and Performance

- **Model Replication** We implement some of the benchmark models reported in the original paper, including OLS + H, OLS-3, OLS-3 + H, Ridge + H, Lasso + H, Elastic Net + H, PLS, PCR, Random Forests and Boosted Regression Trees under different settings. Here '+ H' means that the models use the Huber loss function. The scope of hyperparameter tuning is referred to the table in [1]. However, due to the lack of computational resource, we only took a small number of points in the scope for tuning. For tree based models Boosted Regression Trees and Random Forests, we set the number of trees to only 30 which is also due to the lack of computational resource.
- **Performance Evaluation** Out-of-sample predictive R^2_{OOS} is calculated as the performance metric, where

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

$\mathcal{T}_3$ indicates that fits are only assessed on the testing subsample, whose data never enter into model estimation or tuning. Following [1], we use 'recursive performance evaluation scheme' to evaluate the results. In specific, we choose the data from year 1987 to 2017 as test data, and take only one test year's data at a time for testing; the data for the 12 years prior to the test year will be used as validation set for hyperparameter tuning; the data for the years before the start of the validation set from 1957 will be used as training set. The average out-of-sample predictive R^2_{OOS} for different models in a range with less volatility has been reported in Table 1. We find that PLS can obtain the best performance, and OLS-3, LASSO, Ridge and ENET all outperform OLS, which illustrates that OLS model suffers from overfitting. Besides, tree based methods Boosted Regression Trees and Random Forests outperform most of the methods based on Linear Regression, although the number of trees is only 30. It is predictable that if we increase the number of trees, they will perform even better and will outperform PCR and PLS, which may be benefited from their nonlinear structure. Figure 1 shows the performance of different models with different test year. Over 30-year out-of-sample test, compare with tree-based methods, the linear models exhibit high volatility.
- **Model Complexity** Figure 2 demonstrates the model's complexity for LASSO + H, ENET + H, PCR and PLS. We can find that models with higher degrees of freedom don't always perform better, which also illustrates that OLS model suffers from overfitting.

Table and Figures for Performance Evaluation

OLS + H	OLS-3	OLS-3 + H	LASSO + H	Ridge + H
-10.3384	-3.9414	-6.4574	-2.5651	-7.5897
ENET + H	PCR	PLS	GBDT + H	RF
-2.2151	-0.8557	3.4177	1.5139	-0.9735

Table 1. Average performance (%) of each model.

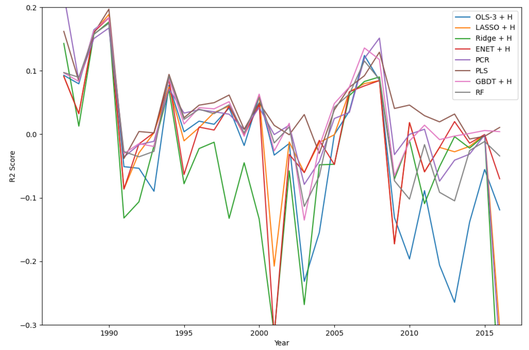

Figure 1. This figure demonstrates performance of different models with different test year.

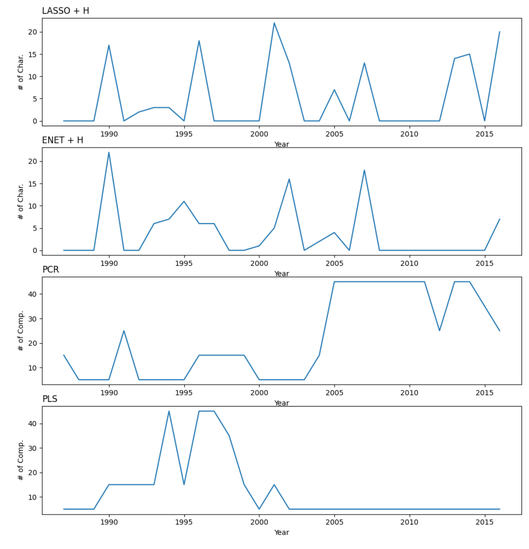

Figure 2. This figure demonstrates the model's complexity for LASSO + H, ENET + H, PCR and PLS in each training sample of our 30-year recursive out-of-sample analysis. For LASSO + H and ENET + H, we report the number of features selected to have nonzero coefficients; for PCR and PLS, we report the number of selected components.

Variable Importance

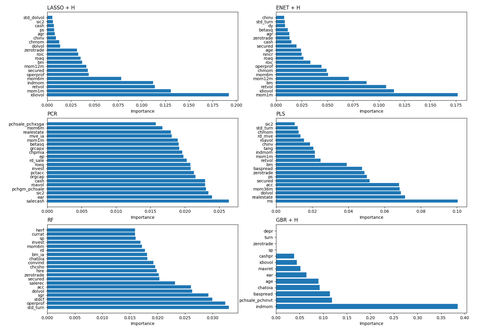

Figure 3. This figure demonstrates variable importance for the top-20 most influential variables in different models. Variable importance within each model is normalized to sum to one. For GBR + H, only 13 variables' importance are nonzero since we only calculate variable importance on top 1000 stocks due to the lack of computational resource.

Following the Section 2.3 in [1], we calculate the reduction in R^2_{OOS} from setting all values of a given predictor to zero within training set as variable importance for the given predictor. Figure 3 reports the resultant importance of the top-20 stock-level characteristics for some methods. Variable importance within the model is normalized to sum to one, allowing for the interpretation of relative importance for that particular model. Due to the lack of computational resource, we only calculate the reduction in R^2_{OOS} within the training set whose corresponding R^2_{OOS} is the best over all the 30 training sets, but not the average. Therefore, the results here are not exactly consistent with those in [1]. However, some of the conclusions in [1] are also reflected here. For example, Figure 3 demonstrates that the Recent Price Trends have great impact (e.g., mom1m, mom12m, chmom, indmom). Besides, Liquidity (e.g., turn, std turn, mvel1, dolvol) and Risk measures (e.g., retvol, idiovol) also have visible impact.

Conclusion and Further Improvement

In this project, we do the paper replication for the paper 'Empirical Asset Pricing via Machine Learning' ([1]). From the perspective of test loss R^2_{OOS}, PLS performs best since OLS suffers from overfitting. Besides, tree-based methods Boosted Regression Trees and Random Forests have potential to outperform linear models if we set more number of trees. For the characteristic importance, Recent price trends, Liquidity and Risk measures have great impact. However, there is still a lot of room for improvement. If we have enough computational resource, we can search for hyperparameters in more detail, increase number of trees for tree-based methods and apply variable importance on all training sets to obtain more convincing results. Besides, neural network method is also worth further exploring. But anyway, by replicating the paper, we gain a deeper understanding of the importance of machine learning in the field of fintech industry.

References

[1] Shihao Gu, Bryan Kelly, and Dacheng Xiu.
Empirical asset pricing via machine learning.
The Review of Financial Studies, 33(5):2223–2273, 2020.

[2] Ivo Welch and Amit Goyal.
A comprehensive look at the empirical performance of equity premium prediction.
The Review of Financial Studies, 21(4):1455–1508, 2008.