MATH 5470 Final Project: Replicate Empirical Asset Pricing via Machine Learning

Qixu Chen [1], Siqi He [2] , Zhiqiu Xia[3] and Meiying Zhang[4] {qchenax, shebb, zxiaae, mzhangcp}@connect.ust.hk
[1,3]: Department of Computer Science and Engineering, HKUST [2]: Department of Operations Management, HKUST
[4]: Department of Electronic and computer engineering, HKUST
Contributions: Qixu: data processing, GBRT and RF models; Zhiqiu: PCR, PLS models; Meiying: Data analysis, Poster; Siqi: OLS and Elastic Net models

1. Introduction

Asset pricing is a fundamental topic in finance, with implications for investment strategies, risk management, and financial decision-making. The original paper [2] introduced a range of machine learning techniques to enhance empirical asset pricing, reflecting the growing interest in applying data-driven approaches to finance. In this study, we aimed to replicate these methods, including linear regression (OLS, elastic net), dimension reduction (PLS, PCR), and tree-based models (gradient boosting trees, random forest). Our objective was to investigate the efficacy of these methods in predicting asset returns and capturing market dynamics, contributing to the ongoing discourse on empirical asset pricing.

2. Dataset

We obtain monthly total individual equity returns from CRSP for all firms listed in the NYSE, AMEX, and NASDAQ. Our sample begins in January 1960 and ends in December 2019, totaling 60 years. We calculate macroeconomic predictors and fill the missing value with cross-sectional median or 0.

3. Data Splitting

➢ **Training Sample:**
This consists of 18 years of data, from 1960 to 1977. Used for training the initial model.
➢ **Validation Sample:**
This consists of 12 years of data, from 1978 to 1989. Used for validating model performance and tuning hyperparameters.
➢ **Out-of-Sample Testing:**
This consists of the remaining 30 years of data, from 1990 to 2019. Used for testing the final model's performance on unseen data.
➢ **Refitting Models:**
Instead of recursively refitting models each month, we refit once every year. At Each time, we increase the training sample by 1 year and maintain the same size for the validation sample by rolling it forward to include the most recent 12 months.

4. Methodology

Below are the algorithms for each of the machine learning methods utilized in our project.

➢ **Linear Regression (OLS, Elastic Net)**

- Ordinary Least Squares (OLS): OLS is a linear regression method that aims to minimize the sum of squared residuals between the observed and predicted values. This simple linear model is bound to fail in the presence of many predictors.
- Elastic Net: The sparsity of the final model is enforced by appending elastic net penalty to the simple linear model. This penalized linear model can produce suboptimal forecasts when predictors are highly correlated.

➢ **Dimension reduction (PCR, PLS)**

- Principal Component Regression (PCR): PCR is a regression method that first performs principal component analysis (PCA) to reduce the dimensionality of the predictor variables and then applies linear regression on the principal components.
- Partial Least Squares (PLS): Same general approach as PCR to reduce the dimensionality. PLS condense the set of predictors from large dimension to a much smaller number linear combinations of predictors.

➢ **Tree-based models (Boosted regression trees, Random forests)**

- Gradient Boosting Trees: Gradient Boosting Trees is an ensemble learning method that builds a sequence of decision trees, with each tree correcting the errors of its predecessor.
- Random Forest: Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.

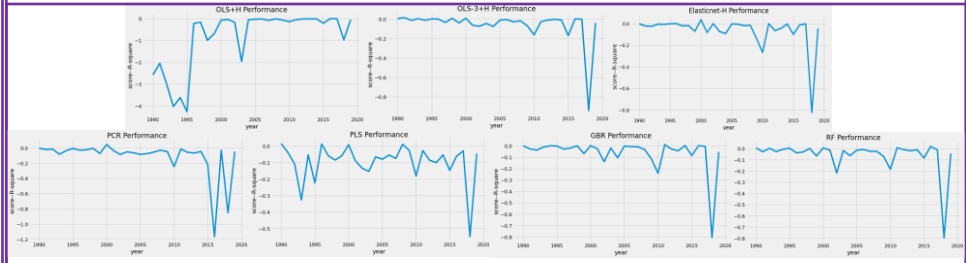
6. Conclusion

In this project, we attempt to replicate the results reported in [2] by implementing various models and comparing their performance and complexities. Consistent with the findings in [2], We observe that tree-based models provide stronger out-of-sample performance while maintaining manageable complexity, and OLS struggles with complexity and performance, likely due to their linear assumptions and sensitivity to high-dimensional data. However, our models' performance did not align with the reported performance, probably due to the different data processing precedure.

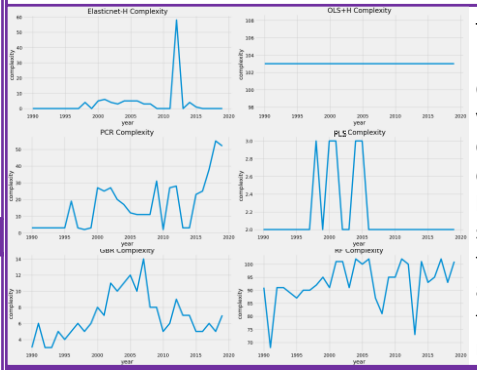
5. Experimental Results and Analysis

We evaluate the models by comparing the predictive performance using the out-of-sample R^2

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



The above figures show the predictive performance of tested models. Among them, OLS has the worst performance, which might be due to the the lack of its regularization compared with OLS-3 and ENet. Tree-based models perform best since they can handle complex nonlinear relationships and provide better resistance to overfitting. We also observe that all models experienced a significant performance drop in 2018. The reason for the drop might be attribute to the unusual market volatility in that year's data and requires further investigation. Overall, there is a gap between our replication and the performance reported in the paper, which is probably due to a difference in our data processing: instead of combining the predictors with each macroeconomic predictors, we directly feed them to the model.



These figures show the models' complexities. PLS has the lowest complexity because it reduces dimensionality by focusing on key latent variables, leading to a simpler model. The high complexity of OLS's may result from the handling of a large number of predictors, especially in our high-dimensional dataset. GBR builds trees sequentially with each tree correcting errors from the previous ones. This sequential approach may be computationally more efficient than RF's parallel approach due to the smaller number of trees.

7. References

[1] Gareth, J. & Trevor, H. (2021). An introduction to Statistical Learning. *Springer, Second Edition*.
[2] Gu, S. & Kelly, B. & and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies* 33(5): 5249-5262.