MATH5470 Final Project
Paper Replication: Empirical Asset Pricing via Machine Learning

lwubf@connect.ust.hk
Department of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST)

Youtube Link:https://youtu.be/MLjOO-qXolg

Introduction

Traditional asset pricing methods have potentially severe limitations that more advanced statistical tools in machine learning can help overcome. In this project, we reproduce some of the work in Gu et al. (2020), in which they provide a comprehensive analysis of expected returns using machine learning techniques.

In this project, we implement the work of "Empirical Asset Price via Machine Learning". After data cleaning, a set of models, e.g., OLS, ENET, PLS, PCR, RF, and NN5, etc., are constructed to obtain the predictive results. We calculate R_{oos}^2 to measure the performance of each model. Furthermore, we report the resultant importance of the top-20 stock-level characteristics.

Dateset

(1)We construct several macroeconomic predictors that are not provided in the original dataset.

(2)We handle the missing values by two strategies: (a) If the percentage of missing value < 50%, we replace the it with the average value. (b) If the percentage of missing value > 50%, we replace the missing value with the mode value.

(3)We perform standard normalization with sample mean and deviation to facilitate training.

Implementation

We implement the benchmark models reported in the original paper, including OLS, ENET, PLS, PCR, RF, and NN. R_{oos}^2 is calculated as the performance metric.

OLS	-6.65	**NN1**	0.27
Elastic Net	0.12	**NN2**	0.29
PLS	-1.23	**NN3**	0.28
PCR	3.38	**NN4**	0.32
RF	-0.68	**NN5**	0.25

Results and Analysis

According to the results reported in Table, we can conclude that:

1. nn4 can obtain the best performance, we think it is benefited by the nonlinear structure.

2. Compared with the methods based on Linear Regression and tree, the performances of the Non-linear method are more promotive.

The most powerful predictors are associated with price trends, including return reversal and momentum, followed by measures of stock liquidity, stock volatility, and valuation ratios.

Controduction

In this project, we do the paper replication for the problem of empirical asset pricing via machine learning. In summary, for the R_{oos}^2 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.

References:

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273.

Contributions: Linshan Wu (20973459)