MAFS 6010Z Project 2: Paper Replication Study

CHEN Yuying, LOU Ruoyu and SHANG Zhiheng {ychenhn, rlou, zshangaa}@connect.ust.hk
Department of Mathematics, HKUST

1. Introduction

The aim of this project is to replicate some results of the paper <Empirical Asset Pricing via Machine Learning>, which performed machine learning repertoire to predict the expected return and identify informative predictor variables. Six machine learning methods had been used in this project and analysis on prediction and variable importance was conducted thereafter. Although the result of replication is quite different from the original paper, there does have some interesting contents that deserves further discussion.

2. Data Cleaning

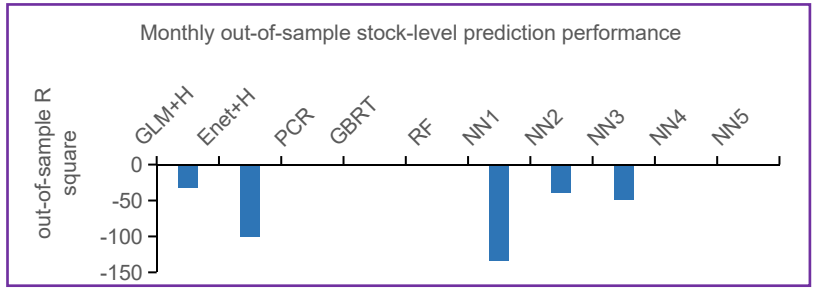
The original data was extracted from the author's website, which was the monthly total individual equity returns from CRSP for all firms listed in the NYSE, AMEX, and NASDAQ. In order to better conduct our replication, we applied one-hot process on Industry variables, replaced NA data with the average value on the same day and randomly chose 500 stock sample for further analysis.

3. Evaluation Method and Prediction Analysis

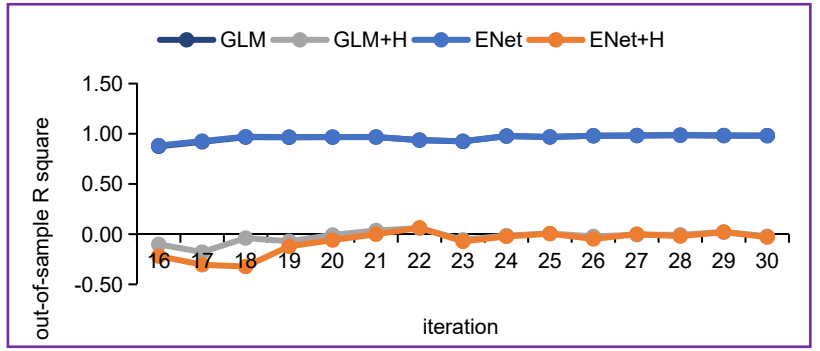
A recursive method was deployed in the evaluation. Through splitting dataset into three parts (training, validation and testing), model refitting could be avoided. In this project we selected 6 methods to conduct the research. Linear regression (elastic net), generalized linear model (ridge), dimension reduction (PCR), trees (gradient boosting trees, random forest) and neural networks was involved. We chose a parameter range, and find the parameter with the smallest error in this range through iteration. The result is shown below.

	GLM+H	Enet+H	PCR	GBRT	RF
R2_oos	-32.5597	-100.579	-2.3E-05	-0.54108	-0.01813
	NN1	NN2	NN3	NN4	NN5
R2_oos	-133.599	-39.9254	-49.7925	-0.00239	-0.00378

Out-of-sample Stock-Level Prediction Performance



From the result above we find that generalized linear model with group ridge (GLM) and elastic net (ENet) perform poorly using Huber loss. As to the neural network architectures with one to five layers (NN1,...,NN5), the out-of-sample R square result converges with the increase number of layers.
In order to find out the effect of Huber loss on linear model, the recursive prediction result is shown below.

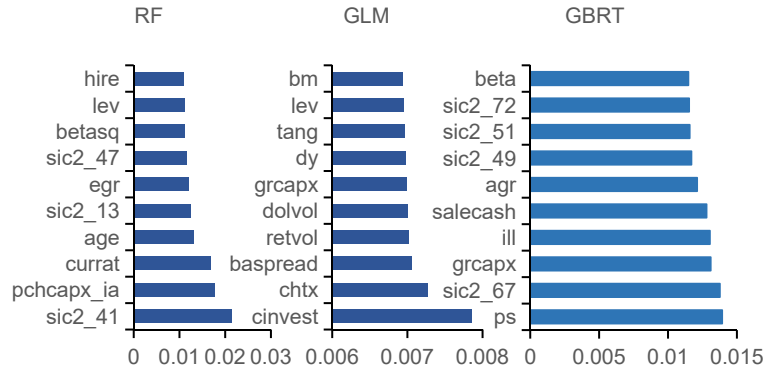


Prediction Performance Comparison Huber Loss Effect

From the picture, in the second half of the recursive calculation, GLM and ENet shows similar pattern and the value of out-of-sample R square converge to 1. While the result of those two linear models under Huber loss converge to 0 instead.

4. Variable Importance

The relative importance of individual covariates for the different model is investigated. The top 10 rankings of characteristics for several models is shown below. The characteristics ranking changes with models but we can still find that sic variables performs well in tree models.



AUC Curve for LGBM

5. Conclusion

The reason behind the pattern conflict may owe to several factors. The dataset chosen in this project is limited due to the gap of calculation power. Considering the conservative nature of Huber loss, the performance is restrained by the loss function is also possible.

6. Contribution

Replication for PCR, GLM , Elastic Net and variable importance
CHEN Yuying 20744353
Replication for Tree Models and Neural Networks
LOU Ruoyu 20743763
Data Analysis and Report Drafting
SHANG Zhiheng 20738938