Project 2: Paper Replication Study of *Empirical Asset Pricing via Machine Learning*

Team Member: Fu Qiyin, Liu Enjie, Yu Xintong, Zhao Encong {qfuab, ezhaoab, eliuac,xyubi} @connect.ust.hk

1. Introduction

The objective of this project is to do paper replication study of *Empirical Asset Pricing via Machine Learning* and to measure asset risk premiums via machine learning. Six models were used in the work which are **Elastic Net, PCR, PLS, RF, GBRT and NN**, and R2 were calculated for their performance. Feature engineering were conducted by data cleaning and variable interaction calculation. For each model, we calculated the reduction in R2 from dropping a given predictor to get characteristic importance and divided the top-20 most influential variables four categories and do some analysis.

2. Data Cleaning & Missing Value

According to the paper, we eliminated variables which do not appear in the reference. For nominal variables, such as sci2, they were re-coded by onehot encoding. Since data without price were considered to be useless, they were drop directly and the remaining missing value were filled with 0.

3. Variable Interaction

3.1 Extra consideration of interaction for some models

Some models themselves consider interaction between features such as tree models -- GBRT, RF, etc. In this case, we can directly use the processed dataset. But for other models, such as Elastic Net, PCR, PLS, etc, we manually calculated the interactions between variables and used them as new features.

3.2 How we calculated interaction?

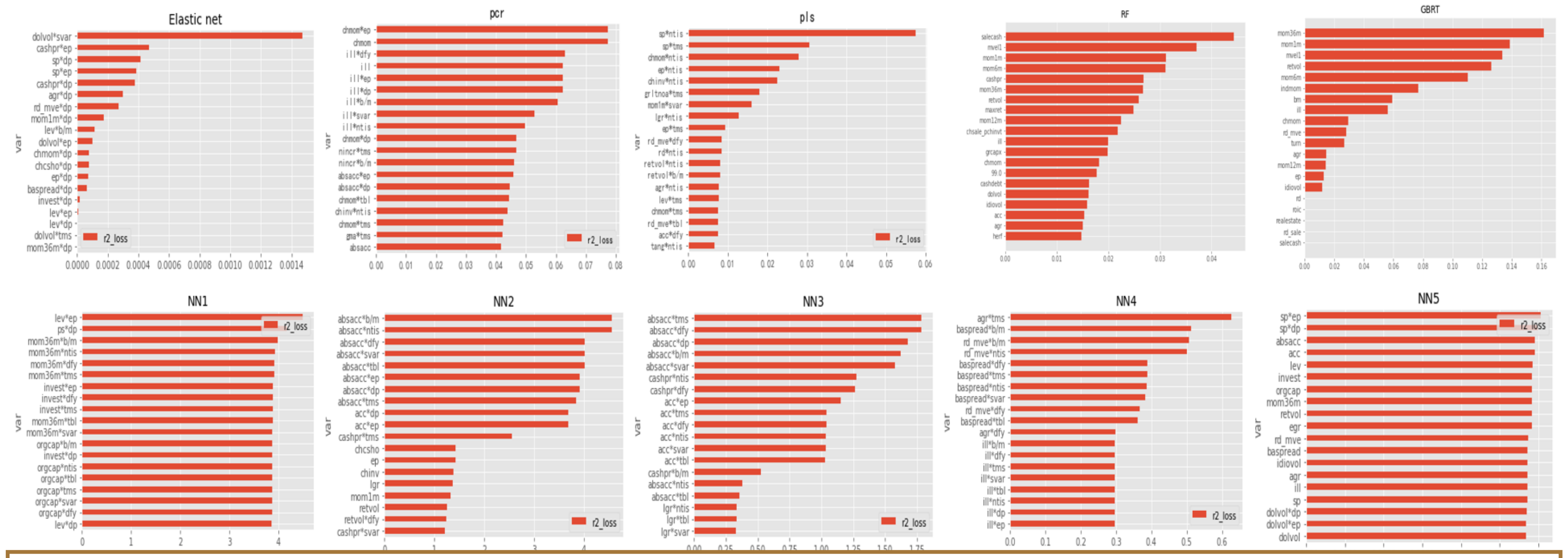
In the case of the original dataset, we need to create interaction features of 94 stock features and 8 macroeconomic features. Considering computer memory and operating speed, we made some adjustments by selecting features at first and then creating interactive features. **Pearson correlation coefficient** is applied in feature selection part. By multipliying **30 selected** stock features and 8 macroeconomic features, we get 240 new variables. The total 270 features then were used in model.

4. Model Performance

model	R²	model	R²
PLS	-0.274	NN1	-39.996
PCR	-0.058	NN2	-19.373
Elastic Net	0.00299	NN3	-1.381
RF	0.00378	NN4	-0.833
GBRT	0.00391	NN5	0.0035

According to the **recursive performance evaluation scheme**, we calculated **the average R2 of 30 batches** to get the final performance of each model. Results are shown in the left table. The best models are GBRT, RF, NN5, which are slightly different from the result of paper. We consider: (1) different processing and calculations of the data are thought to be a cause. (2)another reason may be that we did features selection, which is not mentioned in the paper, either.

5. Variable Importance



By removing one variable at a time and calculating the reduction R2, we can get the performance of each variable. The more reduction in R2, the more important the corresponding variable is. Results are shown in the above charts. We can see **interaction variables** play well in many models, which indicates that the interaction of stock features with macroeconomic factors are significant.

We also analyze the **stock-level features**. The most influential stock-level predictors can be divided into 4 categories .

① Momentum: short-term reversal (mom1m), stock momentum (mom12m), momentum change (chmom), industry momentum (indmom), recent maximum return (maxret), and long-term re versal (mom36m).

② Liquidity, including turnover and turnover volatility (turn, SD_turn), log market equity (mvel1), dollar volume (dolvol), Amihud illiquidity (ill).

③ Risk: total and idiosyncratic return volatility (retvol, idiovol).

④ Fundamental: earnings-to-price (ep), asset growth (agr).

6. Conclusion

In this project, we do paper replication study anf used maching learning method for measuring asset risk premiums problem. For model performance, **GBRT, RF and NN5** are top 3 models according to their testing R2. And **4 types features** (momentum, liquidity, risk, fundamental) and **interactive features** are both the most influential for forecast accuracy.

7. Contribution

Coding: Fu Qiyin, Liu Enjie, Yu Xintong **Poster:** Zhao Encong, Liu Enjie