Paper Replication: Empirical Asset Pricing via Machine Learning

MATH 5470 Final project

Group member:

CAO, Zhefeng

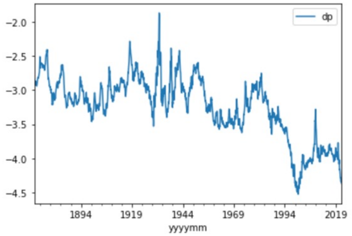
CHU, Mengyuan

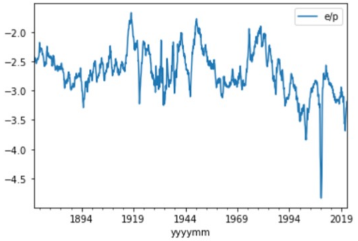
Problem Background

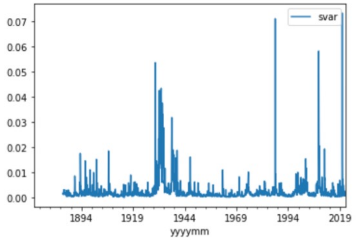
- Empirical Asset Pricing problem
- The definition of the risk premium(**excess return**)
- Predict the **risk premium from some informative predictors of companies and the whole market.**
- Regression problem
- Introduce the predictors **composition**

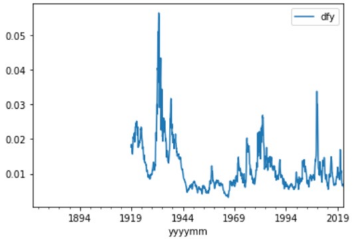
Dataset

- 94 stock-level predictive characteristics
- 8 macroeconomic predictors









- Company type represented by the first two digits of Standard Industrial Classification (SIC) codes
- Monthly data from 1957 to 2016
- The number of stocks: almost 30,000, average 6200

Composition of input

- Because of the cross-section of stock return, we focus on the influence of characteristics for each stock, and the comprehensive impact with the market.

$$z_{i,t} = x_t \otimes c_{i,t}$$

Where ci,t is the characteristics for each stock I at time t, x_t is the macroeconomic predictors at time t

Methodology

- Regression mathematical model

$$r_{i,t+1} = g(z_{i,t}) + \epsilon_{i,t+1}$$

Where r_i,t+1 is the excess return of stock I at time t+1, g is the estimated model, and \epsilon is the residual

OLS

- Linear model

$$g(z_{i,t};\theta) = z_{i,t}^T\theta$$

- Object function

$$\min_{\theta} \quad \mathcal{L} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(r_{i,t+1} - g(z_{i,t};\theta))^2$$

Elastic Net

$$\min_{\theta} \quad \mathcal{L}(\theta) + \phi(\theta; \cdot)$$

where

$$\phi(\theta; \lambda, \rho) = \lambda(1-\rho)\sum_{j=1}^{p} |\theta_j| + \frac{1}{2}\lambda\rho\sum_{j=1}^{P}\theta_j^2$$

PCR & PLS

- Dimension Reduction: Reduce dimension p(920) to small one(10-100)

$$g = (z^T \Omega_K)\theta_K$$
$$\Omega_K = [w_1, w_2, \cdots, w_K]$$

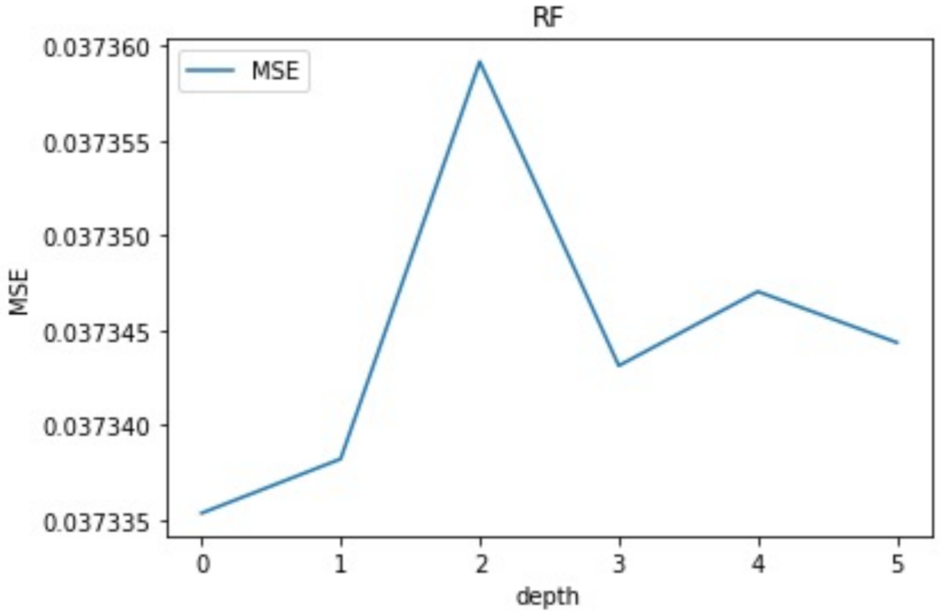
PCR

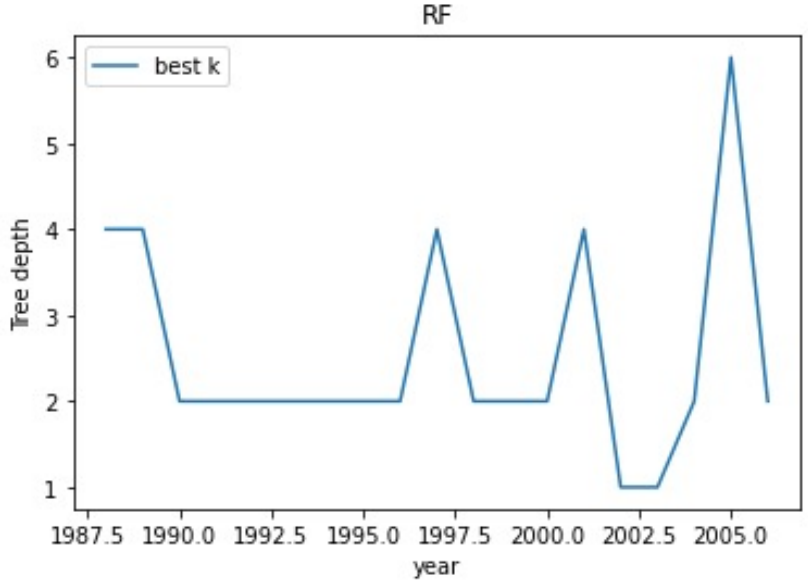
$$w_j = \arg\max_w \quad Var(z^T w)$$
$$s.t. w^T w = 1,$$
$$\mathrm{Cov}(z^T w, z^T w_l) = 0, l = 1, 2, \cdots, j-1$$

PLS

$$w_j = \arg\max_w \quad \mathrm{Cov}^2(R, z^T w)$$
$$s.t. w^T w = 1,$$
$$\mathrm{Cov}(z^T w, z^T w_l) = 0, l = 1, 2, \cdots, j-1$$

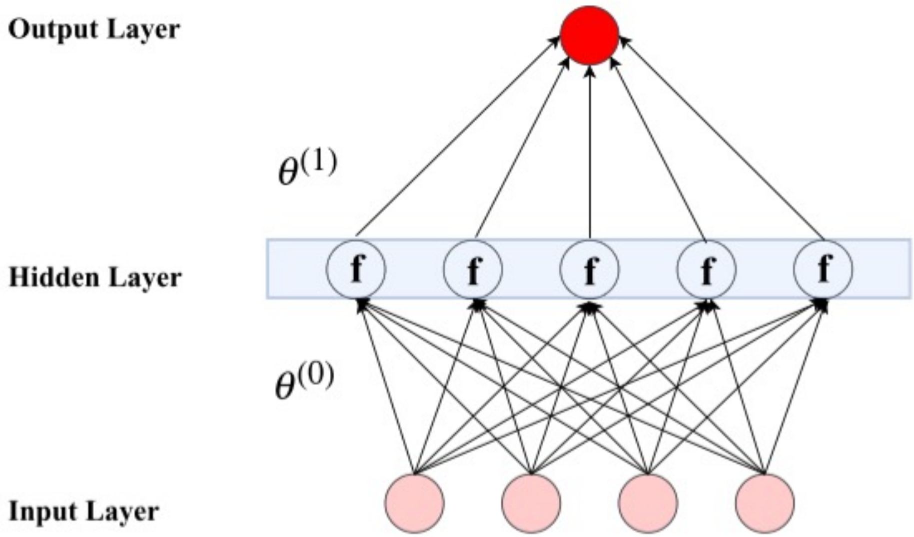
Random Forest





- Depth from 1~6 was tunned
- Feature in each split 3, 5, 10, 20, 30, 50 was tunned

Neural Network

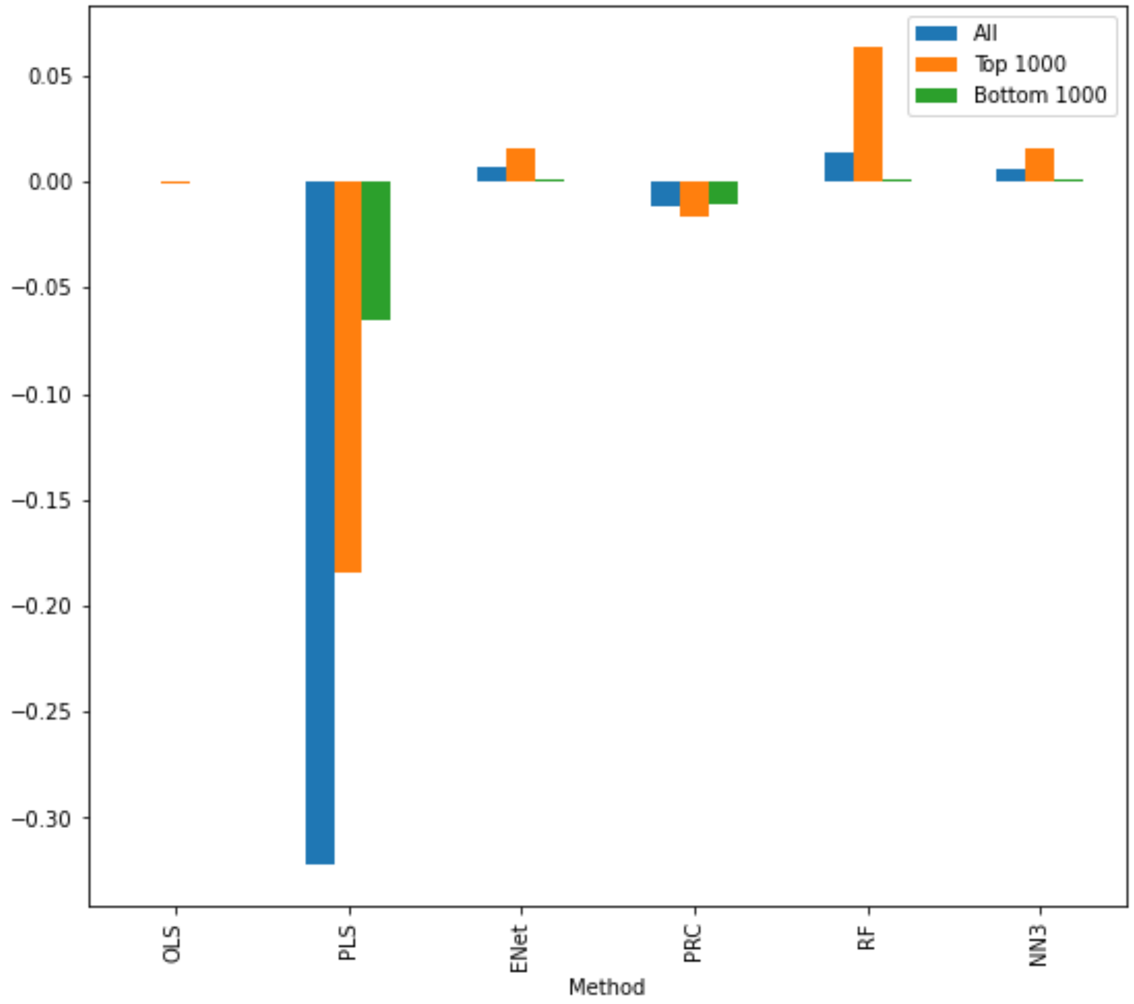


- Three Hidden Layers: 32, 16, 8 neurons
- Nonlinear activation function: ReLU
- Optimizer: Adam SGD
- Batch normalization: batch size 10000

Result Analysis

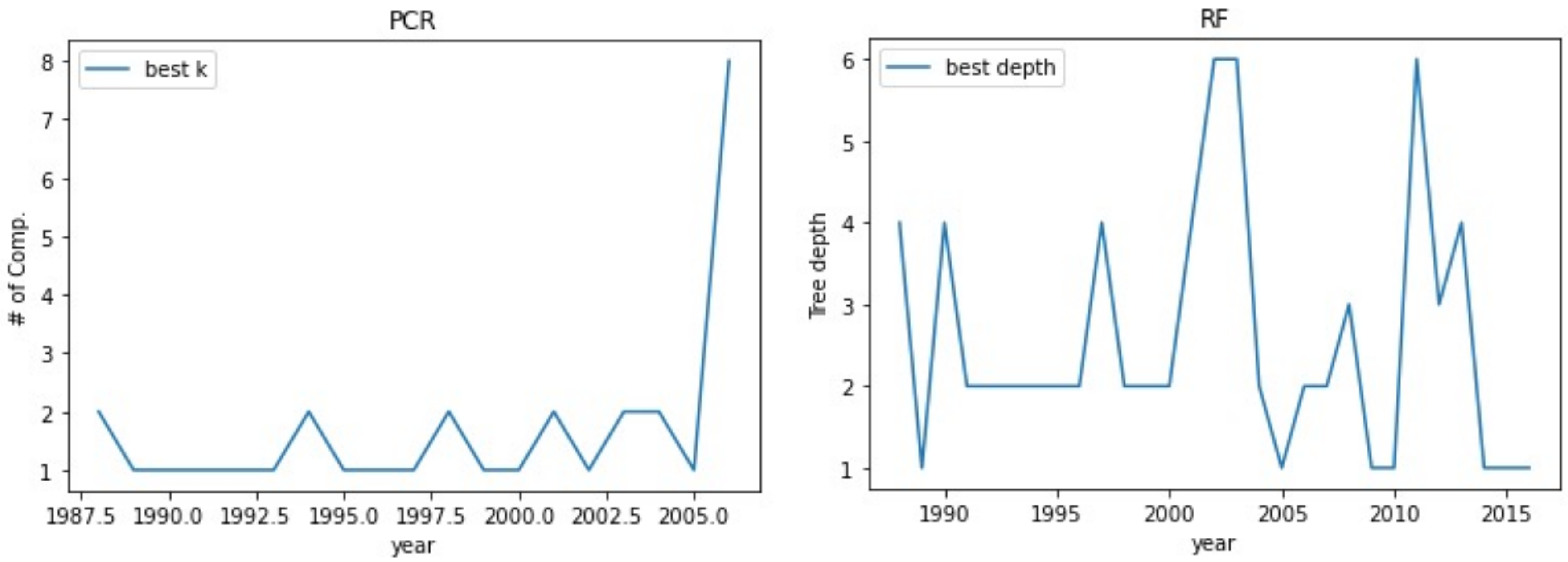
- R2
- DM-test
- Time-varying model complexity
- Variable importance

Coefficient of determination (R2)



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

Time-varying model complexity

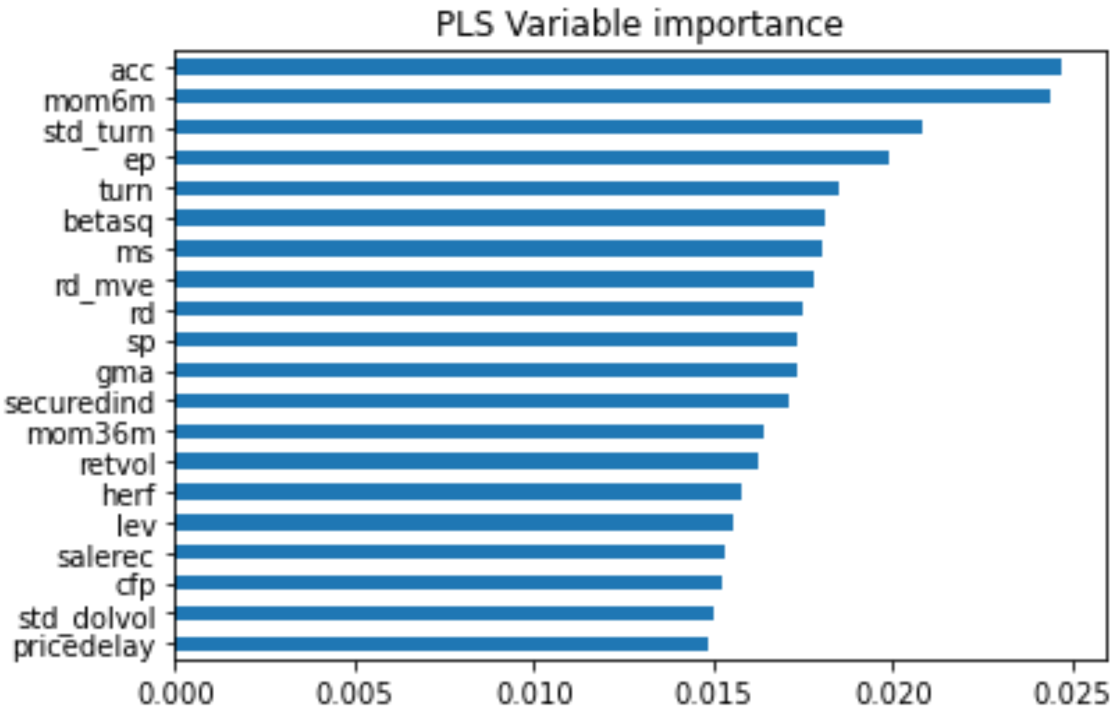


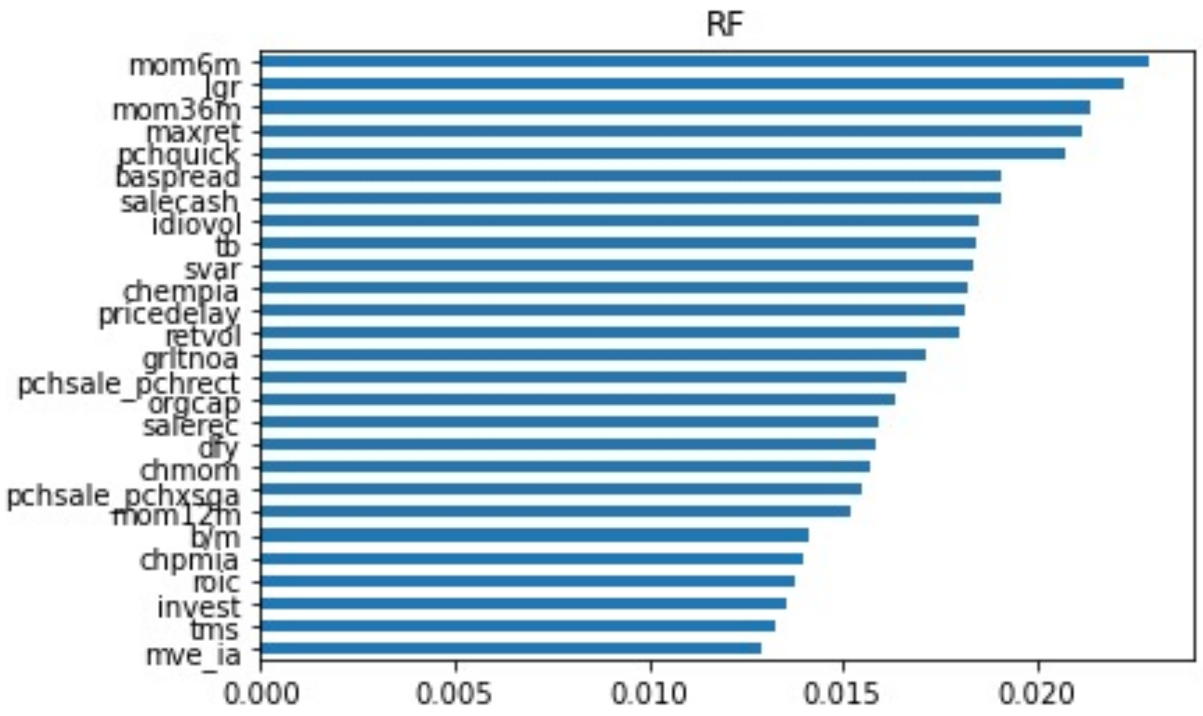
Diebold-Mariano test

	OLS3	PCR	PLS	ENet	RF	NN3
OLS3		-0.455	-1.158	0.202	0.207	-0.272
PCR			-1.120	1.089	0.316	0.213
PLS				1.209	1.174	1.094
ENet					0.077	-0.386
RF						-0.261
NN3						

Diebold-Mariano test statistics comparing the out-of-sample stock-level prediction performance among these models. And positive numbers indicate the column model outperforms the row model.

Variable Importance





Conclusion

- machine learning algorithms can provide important benefit for return prediction
- To gain the robust results, longer period of data are required
- We got better understanding of the whole process from training to testing, especially for high dimensional data
- Hands-on experience dealing with larger data
- improved the coding ability