Project 2:

Replication of the paper “Empirical Asset Pricing via Machine Learning”

Author: Li Shengshu

Abstract

In this project, I chose to replicate the paper “Empirical Asset Pricing via Machine Learning”. For details, the algorithms I chose to replicate are: OLS (including OLS_3), Elastic Net, PLS, PCR, Random Forest, and Neural Network (including NN1, NN2, NN3, NN4, NN5). Furthermore, the train, validation, test datasets are divided by requirement of paper. And the validation datasets are used to find the best value of the changeable parameter in each algorithm. In PLS and PCR, the parameter is the number of main components remained. In Random Forest, it is the depth of the tree. For the rest algorithms, it is the times of iteration of the gradient descent of loss function. Besides, because this replication project is only completed by myself, I mainly tried to replicate the monthly R-squared of each algorithm in their test dataset and the best value of parameters due to limited time and energy.

Data Process

The cleaning of this dataset can be divided in 3 steps.

The first step is “filtering data”, which means that I should only keep the data from 1957/01/31 to 2016/12/31 and the features only mentioned in paper (some unused macroeconomic factors ought to be dropped). Besides, the data will also be filtered by market value, getting top 1000 companies and bottom 1000 companies in each period.

The second step is “filling data”. Since in the dataset above processed, the missing values are so many that if we just roughly fill them as “0” or throw corresponding time series data away, it may well have a bad influence on the final result, I interpolate the sparse missing value by stock code (in dataset, it named “permno”). After those sparse

missing values having been filled, remaining missing values are those appearing in groups, which have high relevant with certain date or data category. Thus, it has less influence on filling those remaining missing values with zero.

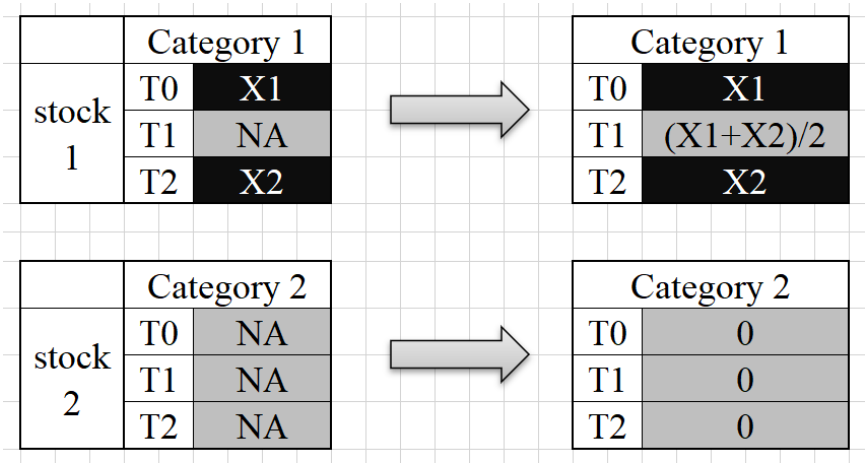

Figure: The sketch map of the "filling data" step.

The final step is "dividing data". For each algorithm, the original dataset needs to be divided into three parts: train, validation, test.

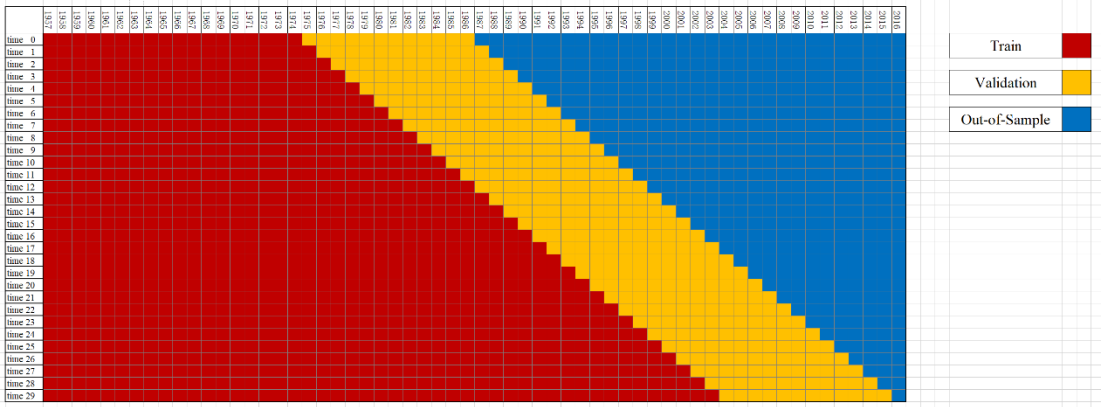

Figure: The sketch map of the "dividing data" step.

Replication

The replication of the R-squared of test dataset

For the convenience of narration, the noun, R-squared of test dataset, will be briefly expressed as R^2_{oos}. Since the classification of train, validation and test dataset is a kind of "rolling", We add a timestamp to each classification. For example, we regard the classification—train: 1957-1974, validation: 1975-1986, test: 1987-2016—as time 0. As the result, we will have 30 classifications with the same number of the R^2_{oos} and the final classification ought to be time 29. Furthermore, the final R^2_{oos} is the mean of the

all R^2_{oos}s from time 0 to time 29.

$$R^2{}_{oos,i} = (1 - \frac{\sum(Real\ Return\_i_{oos} - Predict\ Return\_i_{oos})^2}{\sum(Real\ Return\_i_{oos})^2}$$

$$R^2{}_{oos} = \frac{1}{m} R^2{}_{oos,i}$$

The method to optimize the parameter in each algorithm is finding the smallest MSE in validation dataset. The expression of MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Real\ Return_{validation,i} - Predict\ Return_{validation,i}\right)^2$$

After the definitions were set up, the only thing left to replicate is time. And the detail information for those time consuming algorithms is as follow:

Algorithm	All	Top	Bottom
OLS_3	0.005771	0.064484	0.001334
OLS	-0.096041	-0.027001	-0.045511
PLS	0.002519	-0.019880	0.002181
PLR	0.004439	0.008730	0.001633
Random Forest	0.003343	0.005079	0.002989
Neural Network_1	0.003209	0.005051	0.003889
Neural Network_2	0.003985	0.006372	0.004407
Neural Network_3	0.004013	0.006986	0.005103
Neural Network_4	0.003897	0.006724	0.004012
Neural Network_5	0.003703	0.006431	0.003997

Table: The R^2_{oos} of each algorithm classified by market value of corresponds company's stock.

The replication of find the best parameter of each algorithm

In this replication, the MSE in validation dataset is the only criteria to find the best parameters. In each algorithm, all possible values of the target parameter are fitted in corresponding model, and when MSE takes the minimum value, the corresponding parameters are regarded as the optimal parameters.

$$Parameter\ Value = \arg\min_{x} MSE\left(Algorithm(x)\right)$$

Since different validation datasets can obtain the different optimal parameter, and we do have different validation dataset with corresponding time stramp (from time 0 to time 29), the optimal value of parameter of each algorithm will be a time series.

		T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
PLS	All	1	2	2	2	2	1	1	1	1	1
	Top	1	1	1	1	1	1	1	1	1	1
	Bottom	2	1	1	1	1	6	6	6	7	6
		T10	T11	T12	T13	T14	T15	T16	T17	T18	T19
	All	1	1	1	1	1	1	1	1	3	3
	Top	1	1	1	1	1	1	2	2	2	3
	Bottom	2	5	4	4	5	5	5	5	2	2
		T20	T21	T22	T23	T24	T25	T26	T27	T28	T29
	All	3	2	2	2	3	7	6	4	4	2
	Top	3	3	3	2	3	3	2	3	2	2
	Bottom	3	3	3	3	3	3	3	2	2	2
		T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
PCR	All	5	5	20	20	5	5	5	5	5	5
	Top	10	5	5	5	5	5	5	5	5	5
	Bottom	5	5	10	5	55	5	5	35	35	35
		T10	T11	T12	T13	T14	T15	T16	T17	T18	T19
	All	5	5	5	5	5	5	5	5	5	5
	Top	5	5	5	5	41	10	10	10	10	10
	Bottom	3	45	40	60	95	95	95	95	95	95
		T20	T21	T22	T23	T24	T25	T26	T27	T28	T29
	All	10	5	70	70	70	75	40	40	20	20
	Top	10	10	65	10	65	10	10	5	5	5
	Bottom	95	90	40	90	90	90	90	55	90	90

Table： The number of components for PLS, PCR with minimum MSE in corresponding validation dataset (T0 represent the validation dataset beginning in 1985, rest of each sequence will be postponed for one year until T29 represent 2015)

		T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
RF	All	1	2	1	1	2	4	5	4	5	5
	Top	2	2	1	1	2	4	5	4	5	5
	Bottom	1	2	1	1	2	3	5	5	5	4
		T10	T11	T12	T13	T14	T15	T16	T17	T18	T19
	All	5	5	5	5	5	5	3	3	3	3
	Top	1	2	2	2	2	4	5	4	5	5
	Bottom	5	5	5	5	5	5	3	3	3	3
		T20	T21	T22	T23	T24	T25	T26	T27	T28	T29
	All	3	1	1	1	2	4	4	4	4	4
	Top	1	2	2	2	2	3	5	4	6	5
	Bottom	1	2	3	2	2	4	5	4	5	5

Table： The max depth of Random Forest with minimum MSE in corresponding validation dataset (T0 represent the validation dataset beginning in 1985, rest of each sequence will be postponed for one year until T29 represent 2015)

Summary and reflection

From the result of the replication, a conclusion can be found in the second part, finding the optimal value of the target parameters, those reproduced values are very close to the original values, but in the second part, finding the R^2_{oos}s, the reproduced values of R^2_{oos} are not very close to the original values. From my prospective, this inconsistent might cause by different method off cleaning data. This author of the paper might have better techniques to process the dataset.

Since all things are charged by myself individually, I cannot replicate the whole paper. On contrast, I only replicated several main algorithms and counted their monthly R^2_{oos} and times series of optimal values of target parameter of each algorithm. Thus, the whole replication might be rough. For example, the initial dataset could have been cleaned more technical, because the industry average values of company annually values could have been used to fill the missing values, instead of simply filling 0. Furthermore, more parameters, instead of single parameter ought to be used in optimizing the MSE in validation dataset. However, due to the computing power of my computer and my personal time and energy, this replication paper is my best effort. I believe that, if I have another chance with better preparation in the future, I can replicate better.