Paper Replication:
Empirical Asset Pricing via Machine Learning

Group members:
MA Rongyue 20826086
NI Xiaohan 20825846
Peng Junkai 20756772
YE Mengxiang 20799762

November 14th, 2021

Abstract

Measuring asset risk premium is one of the rudimentary parts of empirical asset pricing. There is a trend to synthesize Machine Learning with the financial field. The high-dimension nature of Machine Learning can improve flexibility, especially for complex and big data.

Our project is to replication the paper – Empirical Asset Pricing via Machine Learning, which aims to predict assets' excess return by several machine learning algorithms. In this report, we would like to share our thinking during the replicating process. The report consists of several parts – Data Preprocessing, Models with Different Techniques, and Recursive Evaluation Method.

The author has tested several methods, such as Linear model, Dimension reduction, Generalized linear model, Tree model and Neural network. For our project, we choose to apply OLS-3, ENet, PLS, PCR, GBRT and RF to replicate the result of the paper and identify the best-performing method by comparing out-of-sample R^2.

The take-aways from the project are twofold.

1. We learn to deal with large quantities of data. The dataset contains all stocks listed in the NYSE, AMEX, and NASDAQ, resulting in approximately 30,000 stocks. Furthermore, the duration scale is also quite long, which is 60 years in total.

2. Through trying different techniques and conducting a comparative analysis, we gain a general understanding of the pros and cons of various algorithms, which deepens our understanding of class materials.

Going forward, we can apply other methods mentioned in the paper, such as Neural Network to improve the final score of prediction.

1 Data preprocessing

Before examining different models' comparative analysis, the crucial preliminary steps are to conduct data preprocessing. Admittedly, during the project, this is also the most time-consuming part for us.

The collection of predictor candidates is relatively large. The authors provided us with many predictors that appeared in the research literature to be regarded as having forecasting power.

We have two sources of predictors. The first excel contains hundreds of stock-level predictive characteristics. And the second excel contains dozens of macroeconomic predictors of the aggregate market.

For Missing values in the first excel file, as the paper indicates, we dealt with them by the cross-sectional median.

The methodology of eight macroeconomic predictors is derived from Welch and Goyal (2008). We firstly conducted Backward-fill to fill NaN values because of the continuous behaviours of macroeconomic data. By their definitions, we calculated those parameters needed from the data.They are dividend-pricing ratio (dp), earnings-price ratio (ep), book-to-market ratio (bm), net equity explansion (ntis), Treasure-bill rate(tbl), term spread (dfy) and stock variance (svar). We merged those factors with the stock-level predictive characteristics and resulted in a bigger comprehensive dataset.

The data downloaded have been primarily cleaned. Following, we can process them to make them more suitable and comprehensive for building and training Machine Learning models.

There are several steps for Data Preprocessing:

(1) Handling Null Values

(2) Standardization

(3) Handling Categorical Variables

(4) One-Hot Encoding

2 Models with different techniques

We have conducted five popular Machines Learning techniques. The following parts illustrate the methodology of the collection of Machine Learning methods.

2.1 Simple linear: Ordinary Least Squares

The starting point of our model selection is the most simple one – Ordinary Least Squares, which can accommodate linear relationships.

The OLS-3 method is considered a benchmark. Our model conducts linear regression on three stock level predictors: size, book-to-market, and momentum.

The advantages are undeniable. OLS-3 model is parsimonious and straightforward. Moreover, the model is also highly selected. Those selected characteristics are regarded as one the most popular predictor in the industry practice.

Moreover, our project also implemented the penalization factor – Huber Loss- to avoid overfitting biased and false discovery.

2.2 Penalized linear: Elastic Net

The failure of OLS leads us to approach the machine learning tool kit aids "Elastic Net". Elastic Net combines the L1 and L2 penalties from both the Lasso and Ridge methods to improve the regularization of statistical models.

Through parameter shrinkage and variable selection, Elastic Net can be used to limit the regression's degree of freedom.

The regularization procedures can help deal with overfitting problems. Hyperparameters can control the model complexity. For instance, the penalization parameter in Elastic Net can help to improve out-of-sample R^2.

The choice of penalty function is

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

2.3 Dimension reduction: PCR and PLS

There are two classic ways to reduce dimension introduced in this sector, PCR(principal component regression) and PLS(partial least squares). They both reduce the dimension of predictors to a much smaller number.

2.3.1 PCR

PCR(principal component regression) is to make an orthogonal transformation of the design matrix without constant terms, that is, the new independent variable is the linear combination of the original independent variables. It is an improvement of ordinary least squares estimation, and its parameter estimation is a biased estimation. PCR is a regression analysis method to estimate the regression coefficients, which is unknown in a standard linear regression model based on PCA (principal components analysis).

PCA uses the idea of dimension reduction. In practical operation, orthogonal rotation transformation often reduces the number of independent variables when little information is lost. Less independent variables are often needed, so its role is self-evident.

There are two steps to perform the method: the first step is PCA(principal components analysis) which is applied to the training data to be a kind of regularized procedure and also a type of shrinkage estimator; and then the second step is to train on the transformed samples by a regression.

PCR is normally used to overcome the multi-collinearity problem by excluding some principal components with low variance when following the regression to get the result of dimension reduction.

2.3.2 PLS

Generally speaking, PLS (Partial Least Squares) can be used if principal component analysis can be used. Partial Least Squares integrates the advantages of principal component analysis, canonical correlation analysis and linear regression analysis. In the application of ordinary multiple linear regression, we are often subject to many limitations. The most typical problem is the multiple correlations among independent variables. Furthermore, there are few examples, even less than the dimension of variables, and there are multiple correlations between variables. Partial least squares regression is born to solve these complex problems.

PLS (Partial Least Squares) is not only a regressor but a transformer, similar to PCR method to help reduce dimension. Nevertheless, comparing with PCR, PLS transformation is supervised.

PLS repeats the procedure of orthogonalizing predictors and targets with regard to the previous ones on the orthogonalized data set, aiming at forming more than one predictive component. And this could stop until the desired number appears.

2.4 Tree model

Regression trees have increasing become prevalent techniques in Machine Learning for incorporating multi-way predictor interactions.Tree model is the most widely used model in the field of machine learning in addition to deep learning. It is also a model with many variants. The advantage of tree model is that it is easy to understand, relatively difficult to overfit, and consumes less resources during training.

2.4.1 Gradient Boosted Regression Trees

GBRT (Gradient Boosted Regression Trees) is a flexible non-parametric statistical learning technique for classification and regression, which is different from traditional regression and is a very popular Machine Learning approach for incorporating multi-way predictor interactions.

The formula shown below indicates how to calculate the prediction value of a tree with K leaves and depth L.

$$g(z_{i,t};\theta,K,L)=\sum_{k=1}^{K}\theta_k \mathbf{1}_{\{z_{i,t}\in C_k(L)\}},$$

We try to consider and combine many weak or simpler learners to get a strong learner. There is no doubt that the weak learners here are the individual decision trees. All the trees are connected in series and each tree tries to minimize the error of the previous tree. Due to this sequential connection, boosting algorithms are usually slow to learn, but also highly accurate. In statistical learning, models that learn slowly perform better.

Basically, observations groups which have similar behaviors are found by trees, and the growth of each tree needs a sequence of steps. And at each step, a new "branch" sorts the data leftover from the preceding step into bins based on one of the predictor variables.

2.4.2 Random Forest

There exists similarities between GBRT and Random Forest. For example, they both not only tend to select trees with few "leaves"(less than six leaves on average), but also combine forecasts from many different trees. However, Random Forest algorithm gets the results based on the predictions of decision trees. It predicts by taking the average of the output from various trees. Increasing the number of trees increases the precision of the outcome. So normally, it is more precise than GBRT.

Random forest is composed of many decision trees, and there is no correlation between different decision trees. Random forests try to get the result of correlation reduction among trees from different bootstrap samples by using a variation on bagging.

When we carry out the classification task, the new input samples enter, and let each decision tree in the forest judge and classify separately. Each decision tree will get its own classification result. Which classification result of the decision tree has the most classification, then the random forest will take this result as the final result.

Model	OLS3+H	PCR	PLS	ENet	RF	GBRT+H
All	-0.154046	-1.483051	3.390748	13.716658	11.087971	46.206583
Top	0.044924	-0.472789	-0.890017	28.608171	11.908887	49.004801
Bottom	-0.070492	-1.682397	-0.162517	15.830734	12.626830	43.874854

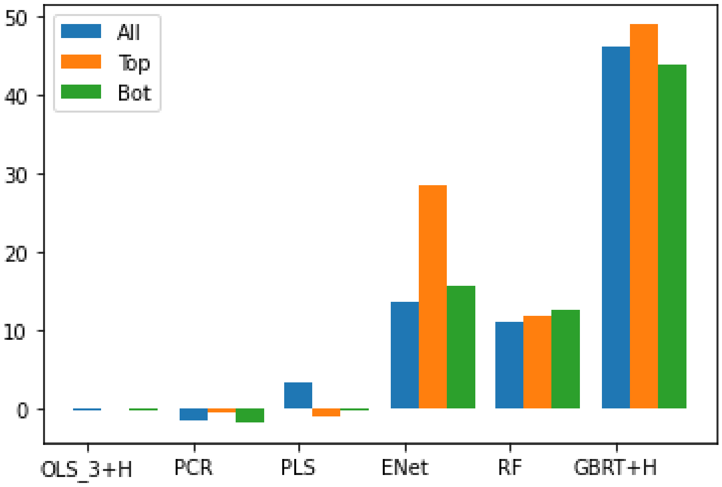
Table 1: out-of-sample stock-level prediction performance (percentage of R$^2_{OOS}$)

3 Individual Stock Returns Prediction

Table 1 presents the comparison of machine learning techniques in terms of their out-of-sample predictive. We compare six models in total, including OLS-3, PCR, PLS, GBRT, and ENet. The first row from Table 1 reports the R^2_{OOS} for the entire pooled sample. The second row in Table 1 is for predictability in large stocks, and the third row is for predictability in small stocks. We could get the R-squared scores from the estimators, which further confirms that GBRT(Gradient Boosted Regression Trees) is the best alternative among the six methods in this case. We could see from the table that there exists negative R-squared. For example, we could find a negative R-squared in PCR method, indicating that PCR performs worse than a regression that would simply predict the mean of the target.

To assess predictive performance for individual excess stock return forecasts, we calculate the out-of-sample R^2.

$$R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}(r_{i,t+1} - \widehat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{i,t+1}},$$

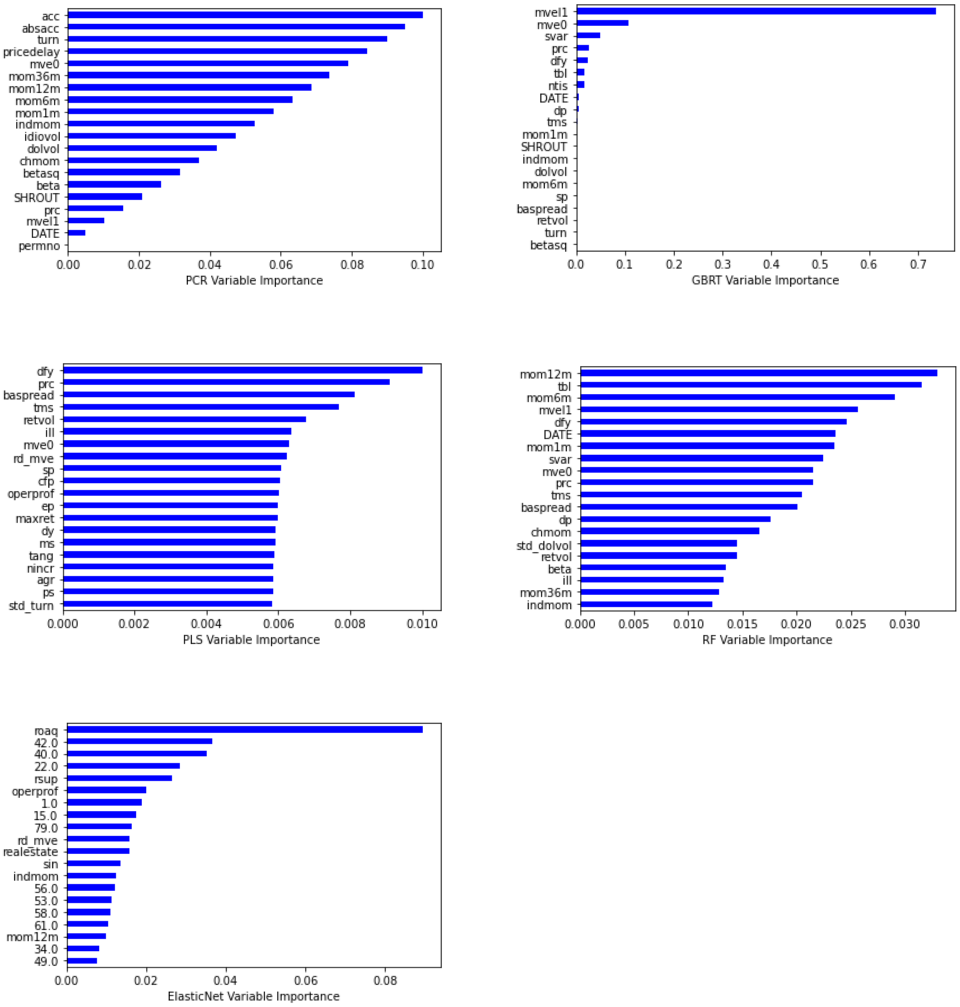


4 Characteristic Importance

Variable importance refers to how much a given model "uses" that variable to make accurate predictions. The more a model relies on a variable to make predictions, the more important it is for the model. It can apply to many different models, each using different metrics.

Modest is the aim for us in machine learning models interpretation. It is vital to identify covariates which have much more significant influence on the cross-section of expected returns, and at the same time, we need to control other predictors here.

This figure shows the resultant importance of the top-20 characteristics for each technique.



5 Reflection

The primary part of our project is a competitive analysis of various Machine Learning techniques. By applying what we have learned in the class to the real financial world is an engaging experience. And this also sheds us lights on that incorporating Machine Learning into empirical financial analysis can improve the overall performance. In a word, Machine Learning are transforming the financial services industry.

References

[1] Gu S , Kelly B T , Xiu D . Empirical Asset Pricing via Machine Learning[J]. Social Science Electronic Publishing, 2018.

[2] Ke Z T , Kelly B T , Xiu D . Predicting Returns With Text Data[J]. NBER Working Papers, 2019.

[3] Welch, I., and A. Goyal. 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21:1455–508.

Group work Contribution

We all actively participated in the initial discussion.

After the outline is settled, PENG Junkai and NI Xiaohan are mainly responsible for coding, and MA Rongyue and YE Mengxiang are mainly responsible for report writing.