math 6010-Zhou-Sun-Huang-Tian Project2 Report

Empirical Asset Pricing via Machine Learning

Zhou Xiaomin 20749212
Sun Ke 20747903
Tian Xinyu 20750015
Huang Yuning 20738524

November 14, 2021

Abstract

In this paper, we use Lasso, Ridge, SGD, PCR, GBRT and random forest methods to predict stock market data through recursive evaluation, and use R^2_{OOS} to evaluate the effectiveness of each method and the importance of each variable.

1 Data

The raw data can be obtained at
https://dachxiu.chicagobooth.edu/download/
and
http://www.hec.unil.ch/agoyal/

1.1 Data Description

The stocks are listed on NYSE, AMEX and NASDAQ, and the timeline is from January 1961 to December 2020. There are 99 variables from 'GKX_20201231.csv' and 8 macroeconomic predictors (dp, ep, bm, ntis,tbl, tms, dfy, svar) following the variable definitions detailed in Welch and Goyal (2008).
Dividend Price Ratio (dp) is the difference between the log of dividends and the log of prices. Earnings Price Ratio (ep) is the difference between the log of earnings and the log of prices. The Book-to-Market Ratio (bm) is the ratio of book value to market value for the Dow Jones Industrial Average. For the months from March to December, this is computed by dividing book value at the end of the previous year by the price at the end of the current month. For the months of January and February, this is computed by dividing book value at the end of two years ago by the price at the end of the current month. Net Equity Expansion (ntis) is the ratio of 12-month moving sums of net issues by NYSE listed stocks divided by the total end-of-year market capitalization of NYSE stocks.

Treasury Bills (tbl) : Treasury-bill rates from 1920 to 1933 are the U.S. Yields On Short-Term United States Securities, Three-Six Month Treasury Notes and Certificates, Three Month Treasury series in the NBER Macrohistory data base. Treasury-bill rates from 1934 to 2005 are the 3-Month Treasury Bill: Secondary Market Rate from the economic research data base at the Federal Reserve Bank at St. Louis.

The Term Spread (tms) is the difference between the long-term yield on government bonds and the Treasury-bill. The Default Yield Spread (dfy) is the difference between BAA and AAA-rated corporate bond yields. Stock Variance (svar) : Stock Variance is computed as sum of squared daily returns on the SδP 500. G. William Schwert provided daily returns from 1871 to 1926; data from 1926 to 2005 are from CRSP.

1.2 Data Processing

1.2.1 Missing Data

We replace all missing values of firm characteristics with 0 and remove the samples with missing returns.

1.2.2 Sample Splitting and Tuning via Validation

We divide our sample into three disjoint time periods that maintain the temporal ordering of the data. The first, or "training," subsample is used to estimate the model subject to a specific set of tuning parameter values. The second, or "validation" sample is used for tuning the hyperparameters. The third, or "testing" subsample is used to evaluate a method's predictive performance.

2 Model Introduction

2.1 Recursive Evaluation

For the 60-year dataset in this article, the first 18 years will be used as the training dataset, the middle 12 years will be used as the validation dataset, and the next 30 years will be used as the testing dataset. After completing a model estimation with the training dataset and the validation dataset, we make predictions on the subsequent year's testing dataset and evaluate the performance. Then, we increase the training dataset by one year, move the validation dataset window back by one year which still maintains the length of 12 years, and perform model estimation again, then make predictions and evaluations on the test set data for the next year. Repeat the above steps until the end.

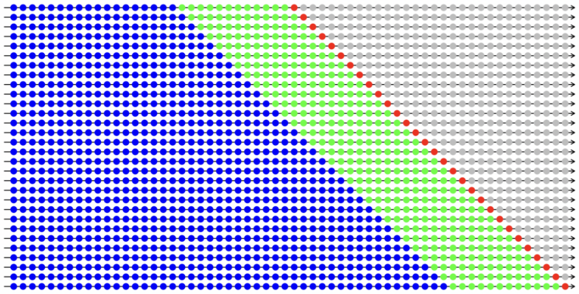


Figure 1: Recursive performance evaluation scheme

2.2 Lasso Regression

LASSO (Least absolute shrinkage and selection operator) is a linear regression method that uses L1-regularization. It constructs a penalty function to make part of the learned feature weights zero, so as to achieve the purpose of sparseness and feature selection, retain the advantages of subset shrinkage. It is a biased estimation for processing data with multi-collinearity.
The characteristic of LASSO regression is to perform variable selection and regularization while fitting a generalized linear model. Regardless of whether the target-dependent variable is continuous, binary, or multiple discrete, it can be modeled and predicted by LASSO regression. The variable selection here refers to selectively putting the variables into the model to get better performance. Complexity adjustment refers to controlling the complexity of the model through a series of parameters to avoid over-fitting. For a linear model, the complexity is directly related to the number of variables in the model. The more the number of variables, the higher the complexity of the model. More variables can often give a seemingly better model when fitting, but it also faces the danger of over-fitting.

2.3 Ridge Regression

Similar to LASSO, Ridge also reduces the complexity of the model and selects variables by adding a penalty function, thereby reducing overfitting. The difference is that ridge uses L2-

regularization to limit the eigenvalues. Ridge Regression tends to evenly distribute the weights among related features, while Lasso tends to select one of the related features, and the weights of the remaining features decay to zero.

2.4 SGD

SGD(Stochastic Gradient Descent) estimates the loss gradient of each sample, and in the process updates the model with a decreasing learning rate. The model we use is the linear model fitted by minimizing a regularized empirical loss with SGD.

The specific steps of this method are: randomly select a group from the sample, update the model parameters once according to the gradient after training, then extract another group, and update the model parameters again. In each iteration process, the samples will be randomly shuffled, effectively reducing the parameter update offset problem caused by the samples. Therefore, in the case of extremely large sample sizes, it is possible to obtain a model with a loss value within an acceptable range without training all samples. When the training data is too large, the gradient boosting method may cause insufficient memory. At this time, SGD can solve this problem.

2.5 PCR

PCR (Principal Component Regression) is an unsupervised learning method that transforms a large number of correlated variables into several new principal components that can represent most of the original numerous variables. These new principal components are not related to each other, at the same time, it can contain a large amount of information that belongs to the original variable, which can effectively reduce the dimensionality of independent variables and eliminate the influence of multicollinearity.

PCR consists of two steps. First, use principal component analysis to form a small combination of principal component variables, so that it can retain the covariance structure of most of the original variables. Then use the small part of variables to do regression. One disadvantage of this method is that the principal component analysis is based on the covariance between the predictor variables, which occurs before the prediction step, and does not consider how these predictors will be related to future dependent variables.

2.6 GBRT

GBRT (gradient boosted regression trees) is a model that integrates many simple trees for comprehensive prediction. This method first uses a simple tree to fit the data, and then uses the second tree to fit the fitting residuals of first tree, then the later tree is always used to fit the fitting residuals of the previous tree, and finally all the trees are integrated to predict the future returns. The advantages of GBRT are: 1. It can handle different types of data. 2. It is very strong in handling abnormal points outside the space (through a strong loss function). The disadvantage of GBRT is that the scalability is not good, because boosting is naturally executed sequentially and is very difficult to parallelize.

2.7 Random Forest

Random forest is similar to GBRT. It is also predicted by combining the results of many trees, but random forest is a bagging method that uses independent trees to fit independent data generating by bootstrap method, and finally integrate all independent trees to predict . This method reduces the correlation of the prediction, and the variance is smaller than that of the ordinary bagging method.

2.8 Performance Evaluation: The Out-of-sample R^2

We use out of sample R^2 to measure the effectiveness of each prediction of each model.The denominator is the sum of squared excess returns without demeaning, it makes up for the shortcomings of other methods that compare forecasts with historical returns in analyzing individual stock returns.

$$R^2_{OOS} = 1 - \frac{\sum_{(i,t)\in\tau_3}(r_{i,t+1} - \widehat{r}_{i,t+1})^2}{\sum_{(i,t)\in\tau_3}(r_{i,t+1})^2}$$

Here stocks are indexed as i=1,...,n and months by t=1,...,T. τ_3 means that fits are only assessed on the testing subsample.

2.9 Variable Importance

In this report, we use R^2_{OOS} to evaluate the importance of variables. We set all the values of a variable to 0 and keep the other variables unchanged to see how much the whole predicted R^2_{OOS} declines.

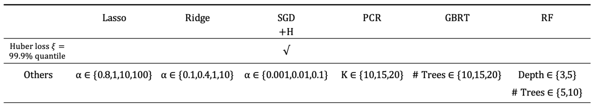
3 Empirical Analysis

3.1 The Comparison of 6 Methods by R^2_{OOS}

We compare each method by comparing the R^2_{OOS} generated by these methods. For each method, our R^2_{OOS} calculation step is divided into three steps: for each data set, we first treat all stock data as panel data and use the data of the training set and the validation set to adjust the parameters and fit the model. Then we predict the data using these fitted models and calculate R^2_{OOSi} (i=1,2,3..) of all stocks for the i-th out of sample data set (we only use the first-year data of each out of sample data set for calculation). Finally, by averaging all these R^2_{OOSi} we have the R^2_{OOS} of this method.

3.1.1 Tuning the Hyperparameters

Tuning parameters are chosen from the validation sample taking into account estimated parameters, but the parameters are estimated from the training data alone.The hyperparameters we tune include, for example, the number of trees in GBRT, the number of random trees in a forest, and also the depth of the trees.Figure1 describes the set of hyperparameters and their potential values used for tuning each machine learning model.



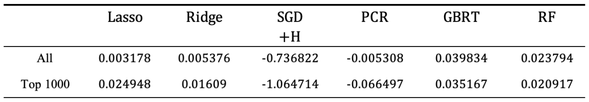
	Lasso	Ridge	SGD +H	PCR	GBRT	RF
Huber loss ξ = 99.9% quantile			√			

Others	α ∈ {0.8,1,10,100}	α ∈ {0.1,0.4,1,10}	α ∈ {0.001,0.01,0.1}	K ∈ {10,15,20}	# Trees ∈ {10,15,20}	Depth ∈ {3,5}
						# Trees ∈ {5,10}

Figure 2: Hyperparameters for all methods

Figure 3 presents the comparison of these 6 methods in terms of their out-of-sample predictive R^2. We compare them using entire pooled samples and subsamples that include only the top 1000 stocks by market equity respectively. For entire pooled sample, the SGD model produces an R^2_{OOS} of -0.74. The reason may be that the SGD model just fit the model using the subsample, though it can solve the insufficient memory problem, it may lead to loss in prediction. It is can be seen that the R^2_{OOS} improves significantly when we use the trees as the model, this is because that lasso and ridge include no interaction among features.

For the subsamples that include only the top 1000 stocks by market equity, lasso and ridge perform better than the first situation, while the other methods perform the same. This indicates that lasso and ridge are more useful in the prediction of large market capitalization stocks.



	Lasso	Ridge	SGD +H	PCR	GBRT	RF
All	0.003178	0.005376	-0.736822	-0.005308	0.039834	0.023794
Top 1000	0.024948	0.01609	-1.064714	-0.066497	0.035167	0.020917

Figure 3: Annual out-of-sample stock-level prediction performance

3.2 Variable Importance

We measure the importance of a single variable by observing the change of R^2_{OOS} after making this variable coefficient zero. The greater the change of R^2_{OOS}, the more important the variable is. This part includes two steps: we first select some variables based on the paper and the correlation between the variables, and then we sort the importance of the variables by observing the changes in R^2_{OOS} when each variable becomes 0.

3.2.1 Variable Correlation

We remove some variables with strong homogeneity and low importance according to the paper and correlation between variables. For the correlation analysis of variables, we first cluster features through K-means and classify highly correlated variables into a class. Then we rank the variables according to the clustering results, calculate the correlation matrix and display it, so that the variables that are highly correlated with each other are ranked together. The following thermal diagram shows the correlation matrix of variables. In the figure, we can refer to variables with strong homogeneity in high-correlation blocks.

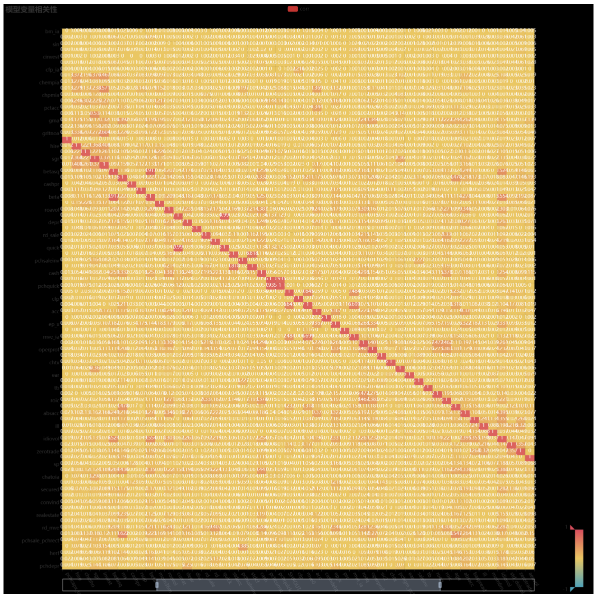


Figure 4: Correlation between features

3.2.2 Variable Importance by Model

We now investigate the relative importance of individual variables for the performance of each model. To begin, for each method, we calculate the reduction in R2 from setting all values of a given predictor to zero within each testing sample, and average these into a single importance measure for each predictor. Figure 5 to Figure 10 report the resultant importance of the top-10 stock-level characteristics for each method. Variable importance within the model is normalized to sum to one, allowing for the interpretation of relative importance for that particular model.

Figure 5 to Figure 10 show that characteristic importance magnitudes for Ridge model and PCR model are highly skewed toward mvel1. SGD model and Trees are more democratic, drawing predictive information from a broader set of characteristics. Lasso model places the same weights on all predictors.

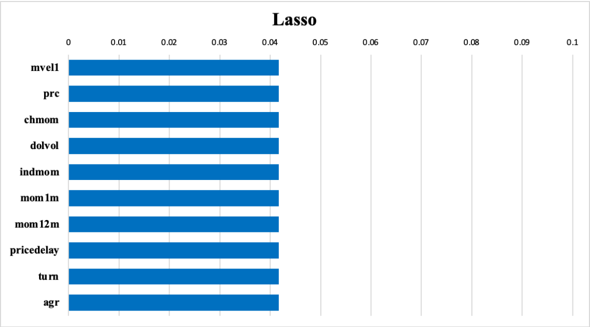


Figure 5: Variable importance by Lasso

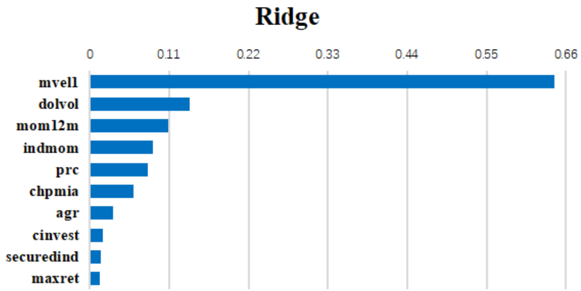


Figure 6: Variable importance by Ridge

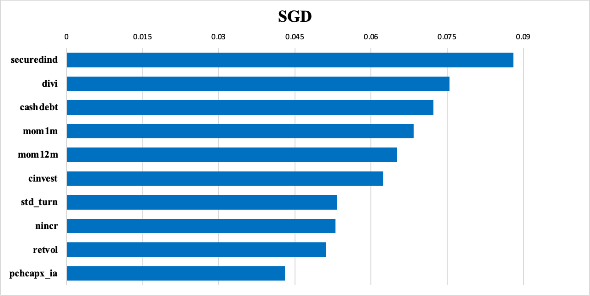


Figure 7: Variable importance by SGD

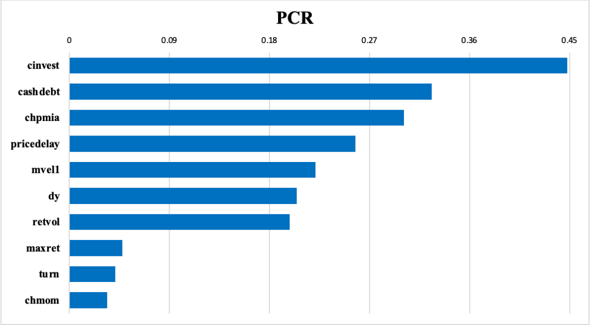


Figure 8: Variable importance by PCR

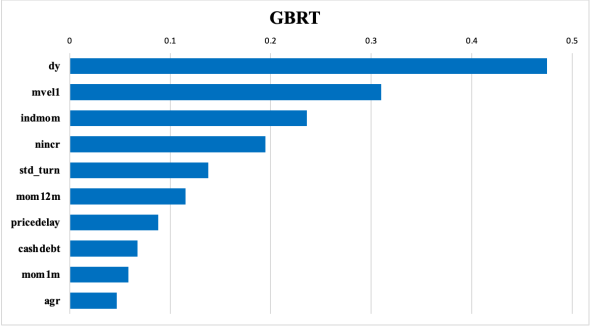


Figure 9: Variable importance by GBRT

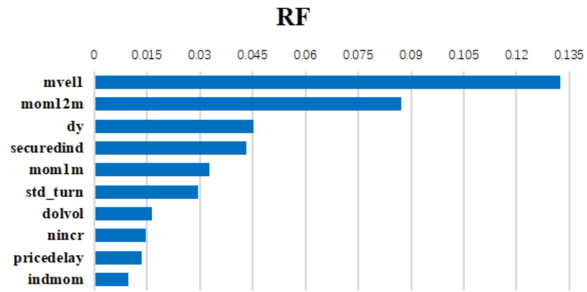
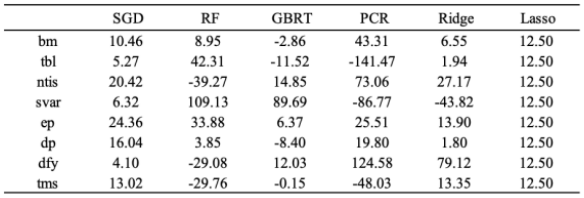


Figure 10: Variable importance by Random Forest



	SGD	RF	GBRT	PCR	Ridge	Lasso
bm	10.46	8.95	-2.86	43.31	6.55	12.50
tbl	5.27	42.31	-11.52	-141.47	1.94	12.50
ntis	20.42	-39.27	14.85	73.06	27.17	12.50
svar	6.32	109.13	89.69	-86.77	-43.82	12.50
ep	24.36	33.88	6.37	25.51	13.90	12.50
dp	16.04	3.85	-8.40	19.80	1.80	12.50
dfy	4.10	-29.08	12.03	124.58	79.12	12.50
tms	13.02	-29.76	-0.15	-48.03	13.35	12.50

Figure 11: Variable importance for macroeconomic predictors

Figure 11 shows the R^2-based importance measure for each macroeconomic predictor variable (again normalized to sum to one within a given model). It can be found that the relative importance of some predictors is negative, which means R^2 increases after we set all the values of them to 0. It is potentially because the model overfits training data. Dividend price ratio (dp) has little role in any model. Lasso places the same weights on all predictors. Other linear models (PCR and Ridge) strongly favor default spread (dfy). Trees place great emphasis on exactly those predictors ignored by linear methods, such as stock variance (svar).

References

[1] Gu, Shihao, Bryan Kelly, and Dacheng Xiu, “Empirical asset pricing via machine learning. Review of Financial Studies,” 33(5), pp. 2223–2273, 2020.

4 Contribution

Huang Yuning: Latex, report
Sun Ke: code, report
Zhou Xiaomin: code, report
Tian Xinyu: code, report