BOSTON UNIVERSITY

Tree Based Methods in Empirical Asset Pricing

Che Wang: Chewangw@bu.edu Xuyang Liu: xyangliu@bu.edu

Advisor: Hao Xing

QUESTROM SCHOOL OF BUSINESS

Introduction

We utilize tree models(Random Forest and GBRT) on SP 500 stocks to implement a long-short strategy to construct machine learning portfolio.
Comparing with SP 500 index, our portfolio achieves a much higher cumulative log-return.

Data

(i) We download the data of SP500 historical components from CRSP and compustat via Wharton Research Data Services.

(ii) Our monthly data begins 1972-01-31, and ends 2020-12-31, 49 year period.

(iii) We split the data in training set, validation set and testing set. The length of the training periods is 14 years, the length of the validation periods is 6 years, and the length of the testing periods is 26 years. We have 4 training/validation/testing periods in total. The start year of each training periods is 1972, 1973, 1974 and 1975.

(iv) We collect 8 macro features that closely related with the market, including stressed-value at risk, earnings-price ratio, etc. And 61 other stock-level features, such as dollar trading volume, industry-adjusted book-to-market ratio, industry sales concentration, net stock issues, number of earnings increases, momentum. We create interaction features by timing each macro features with other features.

Methodology

Gradient Boosting Regression Trees

In order to prevent the overfitting problem common to the decision tree model, the GBRT model combines multiple decision trees. With gradient boosting, decision trees that are shrunken by a factor to preventing overfitting are repeatedly added to fit the prediction residuals from the last tree.
This method is very sensitive to parameters setting.
We set the loss function to be Huber loss and use random search to tune the hyperparameters.
The prediction of a tree, T, with K terminal nodes, and depth L, can be written as:

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

Where the $C_k(L)$ is one of the K partitions of the data. and each partition is a product of up to number L indicator functions of the predictors. θ_k is defined to be the sample average of outcomes within the partition.

Random Forest

Like boosting, a random forest combines forecasts from many different trees. It is a variation on a more general procedure known as bootstrap aggregation, or "bagging". The baseline tree bagging procedure draws different bootstrap samples of the data, fits a separate regression tree to each, then averages their forecasts. Trees for individual bootstrap samples tend to overfit, making their individual predictions inefficiently variable.
Averaging over multiple predictions reduces this variation, and stabilize the trees' predictive performance.

$$H(\theta,C)=\frac{1}{|C|}\sum_{z_{i,t}\in C}(r_{i,t+1}-\theta)^2$$

The procedure is equivalent to finding the branch C that locally minimizes the impurity. Branching halts when the number of leaves or the depth of the tree reach a prespecified threshold that can be selected adaptively using a validation sample.

Performance Evaluation

In this project, we use out-of-sample R^2 to measure the predictive performance.

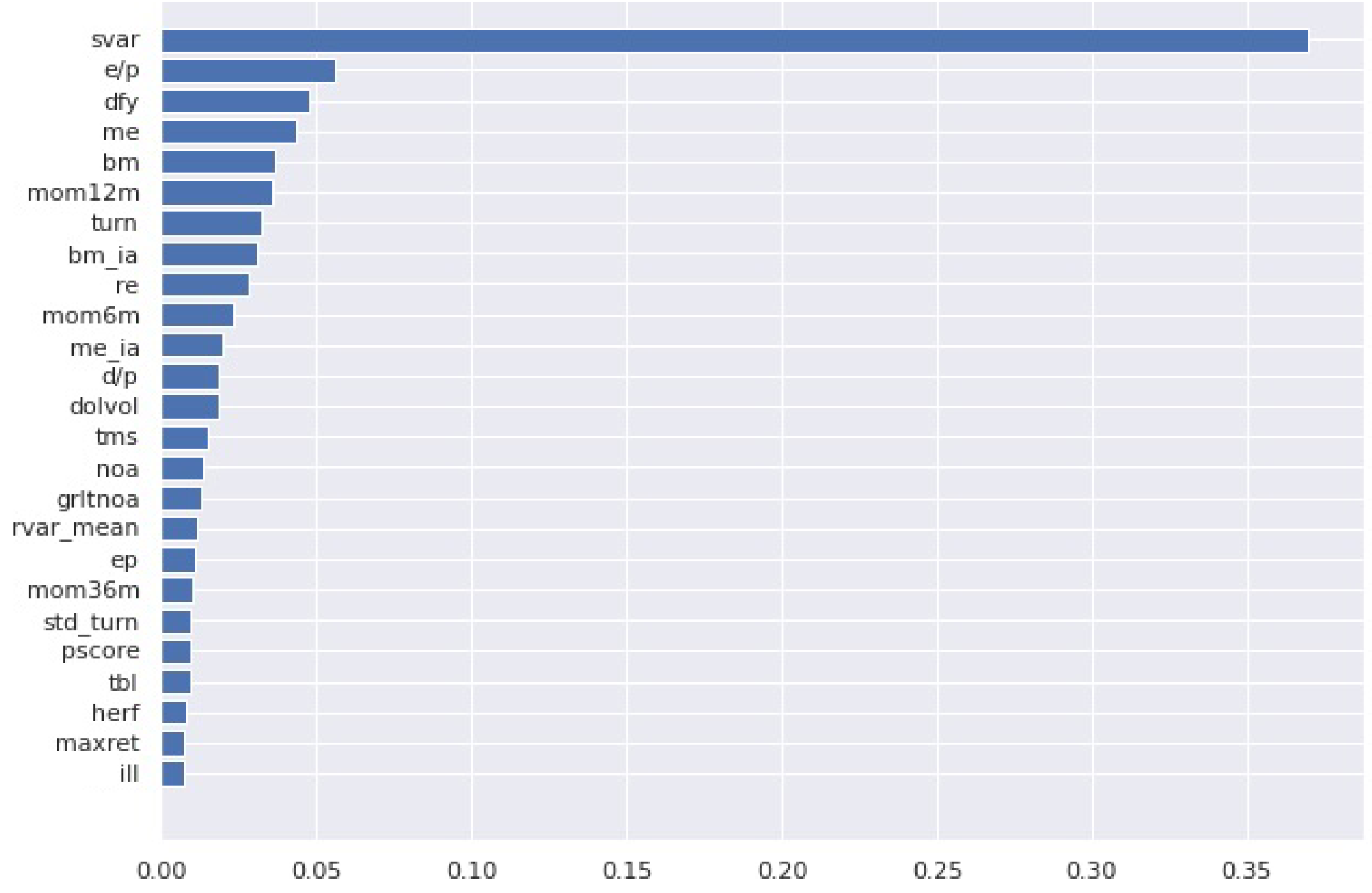
$$R^2_{oos}=1-\frac{\sum_{(i,t)\in T3}(r_{i,t+1}-\hat{r}_{i,t+1})^2}{\sum_{(i,t)\in T3}(r_{i,t+1})^2}$$

(i) As mentioned, we divide the 49 years of data into 14 years of training, 6 years of validation, and 26 years of testing samples. We further roll the whole sample forward to include the most recent 12 months four times to get the average out-of-sample R^2.

(ii) Despite both models' out-of-sample R^2 being less than 5%, our out-of-sample R^2 is higher than similar research. Random Forest is the better performing method based on out-of-sample R^2.

Monthly out of sample stock-level prediction performance (percentage R^2_{oos})

Model	RF	GBRT
R^2_{oos}	4.53	1.43



Feature Importance

Meaning of the first 10 most important stock-level features for random forest model

ABBREVIATION	Meaning	Category
ME	The market equity	Frictions
BM	Book-to-market equity	Value-versus-growth
MOM12M	Momentum rolling 12 months	Momentum
TURN	Shares turnover	Shares turnover
BM_IA	Industry-adjusted book to market	Value-versus-growth
RE	Revisions in analysts earnings forecasts	Intangibles
MOM6M	Momentum rolling 6 months	Frictions
ME_IA	Industry-adjusted size	Frictions
DOLVOL	Dollar trading volume	Trading frictions

To estimate the relative importance of individual covariates, we calculate the reduction in R^2 by setting the specific feature to zero with each training sample. Figure above shows the top 25 features of the RF model. Both models show similar results. Besides macro features, top features can be grouped into stock momentum and liquidity variables.

Value-weighted Portfolio

Reasons for value-weighted portfolio construction:

(i) We choose value-weighted portfolios because stocks with a large market value of equity are often "important" stocks that are traded and held by many people. We want to measure the performance of our models while targeting those important assets.

(ii) The target of our model is to forecast the return of each permno, however, in real life, people often focus on the return of a portfolio rather than single stocks. Thus, we also need to measure the performance of simple portfolios to get a better understanding of the performance of our models.

Details of portfolio construction process:

(i) We construct portfolios for both gbrt and random forest model.

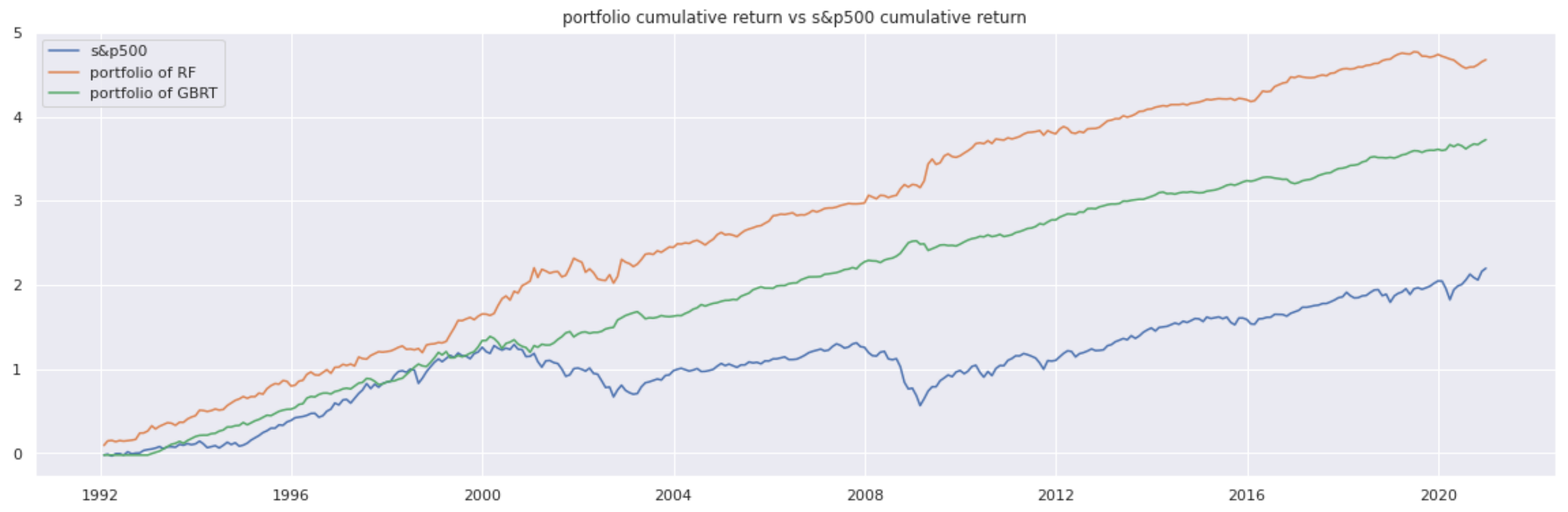
(ii) We set the transaction cost to be 2% and use the S&P500 index as the benchmark.

(iii) Given the models' predicted returns and the market equity of each permno, we first rank the permnos by their predicted returns, then we select the 50 permnos that have the highest average return as the stocks to buy, and the 50 permnos that have the lowest average return as the stocks to sell, and weight each permno according to their market equity.

(iv) Doing so, we get a long-only portfolio and a short-only portfolio for each model, we then combine them to make a long-short portfolio for each model.

Result: the gbrt model and random forest model both have relatively good performance and can beat the benchmark. Most of the time, both portfolios follow the trend of S&P500 index, however, during the period of 2000 to 2003 and the period of 2008 to 2011, the S&P500 index suffers a huge loss, while both portfolios have relatively good performance, showing strong risk resistance capacity.

cumulative return of long-short portfolios



As we can see, the portfolio based on the predicted returns of random forest model has the hightest cumulative return. However, the sharpe ratio(annualized) of random forest's portfolio is around 1.27, which is smaller than the sharpe ratio of gbrt's portfolio, which is around 3.28.

the SR, IR, MDD of each model

MODELS	annualized sharpe ratio	annualized information ratio (with the benchmark of S&P500 index)	maximum drawdown
random forest	3.275392	0.530446	-0.283552
gbrt	1.269521	1.048854	-0.228834

Conclusion

In this project, we try Random Forest and GBRT model on our data and perform a comparative analysis of these methods. Here are some facts of our research:

Our findings show that machine learning would be able to capture the important variables efficiently, which help improve the understanding of assets prices. Both of these tree models perform well and have some important predictive signals in common. The most significant one is the momentum variable.

In addition, the success of our machine learning portfolio construction shows that statistical methods indeed matters in some extent in practical worlds.

Boston university Questrom school of business