Stock Return Prediction via Machine Learning

Wenxuan Ma

Institute of Statistics and Big Data

Renmin University of China

April 24, 2022

Presentation Overview

Motivation

- Measuring asset risk premiums is one of the most canonical problems in asset pricing.

Target

- Compare the predictive performance of traditional statistical models and machine learning methods in the stock return prediction task.

Challenge

- Market efficiency forces return variation to be dominated by unforecastable news.
- Risk premium in stock returns has a low signal-to-noise ratio.

Data Information

Data source

- The monthly total individual equity returns from CRSP.
- The characteristics data are available from Dacheng Xiu's Web site.

Dataset: stocks.csv

- From 1990-01-31 to 2020-12-31, totally 31 years
- The number of stocks is 23,099.
- 2279269 observations, 95 characteristics.

Correlation Analysis

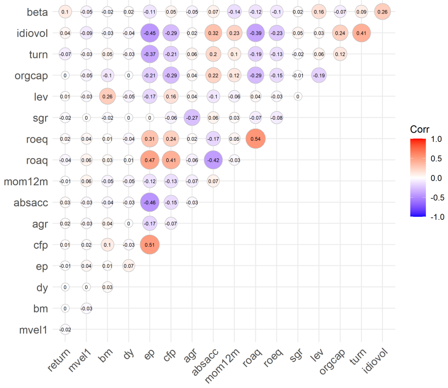


Figure: The correlations between several characteristics and stock return on 2020-12-31

Characteristics Analysis

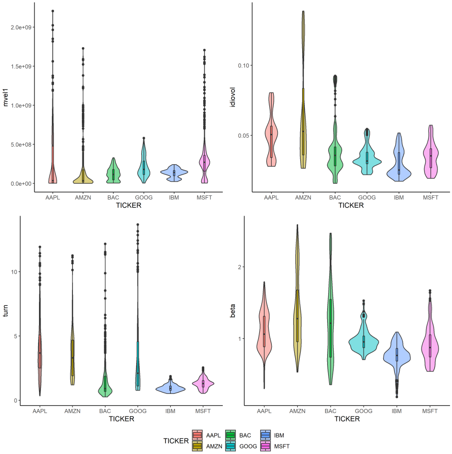
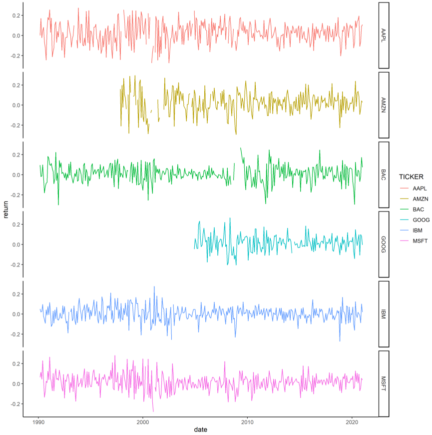
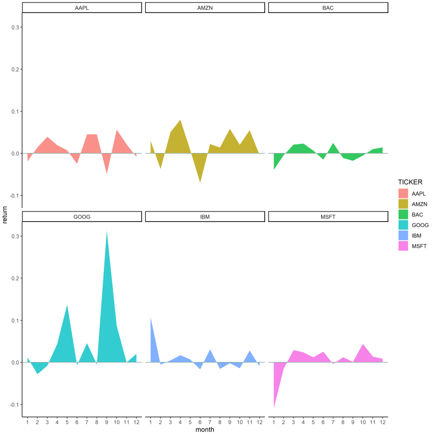


Figure: The distributions of these characteristics are distinct in different stocks.

Changes in Stock Returns



Return time series changes



Average return in each month

Figure: Stock return changing over time

Addictive prediction error model

The general form of the model [Gu et al., 2020] is

$$r_{i,t+1} = \mathbb{E}_t(r_{i,t+1}) + \varepsilon_{i,t+1}, \tag{1}$$

where

$$\mathbb{E}_t(r_{i,t+1}) = g^\star(z_{i,t}). \tag{2}$$

- Stocks are indexed as $i = 1, \cdots, N_t$.
- Times are indexed as $t = 1, \cdots, T$.
- $z_{i,t}$: characteristics.
- Now we assume the function forms of $g^\star(z_{i,t})$.

The Statistical Models

Linear Regression

- $g(z_{i,t};\theta) = z'_{i,t}\theta$
- $\mathcal{L}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(r_{i,t+1} - g(z_{i,t};\theta))^2$

Ridge Regression

- $\mathcal{L}(\theta;\alpha) = \mathcal{L}(\theta) + \phi(\theta;\alpha)$
- $\phi(\theta;\alpha) = \alpha\sum_{j=1}^{P}\theta_j^2$
- $\alpha = [0.1, 1, 10, 100, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000]$
- Optimize the tuning parameters by cross-validation.

Machine Learning: Tree-based Models

Regression Tree (CART)

- The prediction of a tree $\mathcal{T}$ with K leaves and depth L
- $g(z_{i,t}; \theta, K, L) = \sum_{k=1}^{K} \theta_k \mathbf{1}_{\{z_{i,t} \in C_k(L)\}}$

Bagging

- An ensemble method takes the average prediction of the individual trees.
- Random Forest (RF)

Boosting

- An ensemble method takes the sum prediction of a sequence of weak prediction models.
- Gradient Boosted Decision Trees (GBDT), XGBoost

Machine Learning: Neural Networks

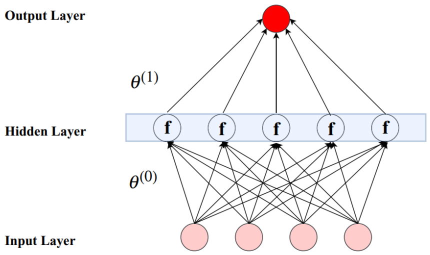


Figure: The neural network with one hidden layer. f is the nonlinear activation function.

- NN1: a single hidden layer of 32 neurons
- NN2: two hidden layers with 32 and 16 neurons
- NN3: three hidden layers with 32, 16, and 8 neurons
- NN4: four hidden layers with 32, 16, 8, and 4 neurons

Model setting

- Activation function:

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{otherwise} \end{cases}$$

- Optimization algorithm: Adam, with learning rate 0.01
- Batch size: 128
- To avoid over fitting: early stopping

A New Tree-based Model

$$\mathbf{r}_t = \sum_{m=1}^{M} \left(g_m^\star (\mathbf{z}_t) + \varepsilon_m \right) \mathbf{1}_{\mathbf{z}_t \in \mathcal{R}_m} \tag{3}$$

- **This model**: identify different regions by splitting uncertainty
- **CART**: aim to decrease the sum of squared errors
- **This model**: the models on the leaf nodes are neural network models
- **CART**: the models on the leaf nodes are constant-valued

Data Splitting and Performance Evaluation

Considering the computation cost, each time we only conduct a five-year prediction and **refit** the model by one year.

- Training dataset (9 years): $2000 - 2009, \cdots, 2004 - 2013$
- Validation dataset (6 years): $2010 - 2015, \cdots, 2014 - 2019$
- Testing dataset (1 year): $2016, \cdots, 2020$

Performance Evaluation:

- RMSE $= \sqrt{\frac{1}{|\mathcal{T}_3|} \sum_{(i,t)\in\mathcal{T}_3} \left(r_{i,t+1} - \widehat{r}_{i,t+1}\right)^2}$
- Out-of-sample $R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3} \left(r_{i,t+1} - \widehat{r}_{i,t+1}\right)^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{i,t+1}}$.

Results - RMSE

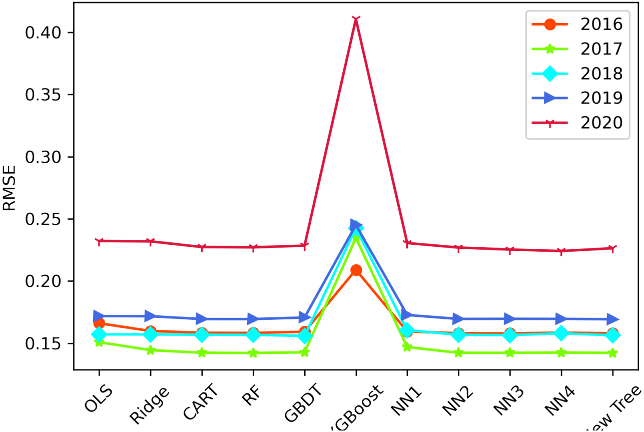


Figure: The predictive RMSE of these methods.

Results - R^2_{oos}

Table: **Yearly percentage** R^2_{oos}: The Bold fonts represent the best performance. The integers in parentheses indicate the rank of models.

Model	2016	2017	2018	2019	2020	Average
OLS	-9.8120(10)	-12.2760(10)	0.9103(8)	-2.8203(9)	-7.4674(10)	-6.218(10)
Ridge	-1.6871(8)	-2.9609(8)	0.9579(7)	-2.7222(8)	-7.2243(9)	-3.5011(8)
CART	0.0868(5)	0.0138(5)	1.2273(6)	-0.0333(3)	-3.0452(6)	-0.8223(6)
RF	0.2431(4)	**0.2811**(1)	1.3832(4)	**-0.0058**(2)	-2.8207(5)	-0.6527(5)
GBDT	-0.9174(7)	-0.4793(7)	**2.3930**(1)	-1.5020(5)	-4.0521(7)	-1.4866(7)
XGBoost	-74.0202(11)	-171.8072(11)	-136.2014(11)	-109.2651(11)	-236.2871(11)	-159.5(11)
NN1	-1.1692(9)	-6.5893(9)	-3.4164(10)	-3.9118(10)	-5.9849(8)	-4.4364(9)
NN2	0.5417(3)	0.0858(4)	1.3164(5)	-0.1403(4)	-2.6560(4)	-0.6099(4)
NN3	**0.7031**(1)	0.1014(3)	1.6758(3)	-0.2408(7)	**-1.2257**(2)	**-0.0533**(1)
NN4	-0.0235(6)	-0.1266(6)	-0.1826(9)	-0.1809(6)	**-0.1456**(1)	**-0.1356**(2)
New Tree	**0.5923**(2)	**0.1583**(2)	**1.8184**(2)	**0.0014**(1)	-2.1647(3)	-0.2877(3)

Summary

Predictive Performance:

- New tree-based model, Random forest and NN3 get the best performance.
- The result of linear regression is bad while adding a penalty function can improve the outcome.
- Surprisingly, the boosting models are not excellent as expected.

Interpretability:

- The traditional statistical models are interpretable.
- Tree-based models possess interpretability due to the modeling processes.
- Although neural network models have high-quality performance in prediction, they cannot interpret their models.

References

Gu, S., Kelly, B., and Xiu, D. (2020).
Empirical asset pricing via machine learning.
The Review of Financial Studies, 33(5):2223–2273.

He, X., Cong, L. W., Feng, G., and He, J. (2021).
Asset pricing with panel trees under global split criteria.
Available at SSRN 3949463

The End

Questions? Comments?