Random Forest for Stock Market Prediction

APS1052 Project

By: Beini Yuting
Lifu Ween

Short-Term vs Long-Term Investment

Short-term investment

A short-term investment is an investment you expect to hold for 3 years or less, then sell and/or convert to cash.

Long-term investment

Long-term investments are assets that a company intends to hold for more than a year.

Fundamental Analysis vs Technical Analysis

Fundamental Analysis

Fundamental analysts study anything that can affect the security's value, including macroeconomic factors (e.g., economy and industry conditions) and microeconomic factors (e.g., financial conditions and company management).

Technical Analysis

Technical analysis is a trading discipline employed to evaluate investments and identify trading opportunities by analyzing statistical trends gathered from trading activity, such as price movement and volume.

Factor Model

$$r_i = \alpha_i + \sum_{k=1}^{K} \beta_{ik} \cdot f_k + \varepsilon_i$$

r_i	expected return
α_i	intercept
β_{ik}	factor loading
f_k	factor value
ε_i	residual error

Two Factors Model

A paper by Chattopadhyay, Lyle, and Wang (2015) presented an exceedingly simple cross-sectional two-factor model that is derived from fundamental financial principles.

Motivation

Lack of widely-accepted and theoretically-motivated empirical asset pricing model

Return-on-Equity (ROE)

$$ROE_{i,t} = 1 + \frac{X_{i,t}}{Book_{i,t-1}}$$

$X_{i,t}$ net income
(variable *ib* from the fundamentals file)

$Book_{i,t-1}$ lag book value of common equity
(variable *ceq* from the fundamentals file)

Book-to-Market Ratio (bm)

$$bm = \frac{Common\ Shareholders'\ Equity}{Market\ Capitalization}$$

Data to Use

inputData_SPX_200401_201312Select.mat,

fundamentalDataSelect.mat

fundamentalsSelect.mat

Random Forest

- Parallel ensemble

- A forest or ensemble of Decision Trees

What are ensembles?

- A set of machine learning techniques combined together to reduce variance, bias or improve prediction

- Two kinds of ensembles
  - Sequential ensemble (AdaBoost)
  - Parallel ensemble (Random Forest)

Random Forest

A forest of Decision Trees

- Decision tree is a series of “if…then...” questions to predict

- Randomly pick (bootstrap) samples with a subset of bootstrapped features to train the model

Parallel so that each tree won’t affect each other

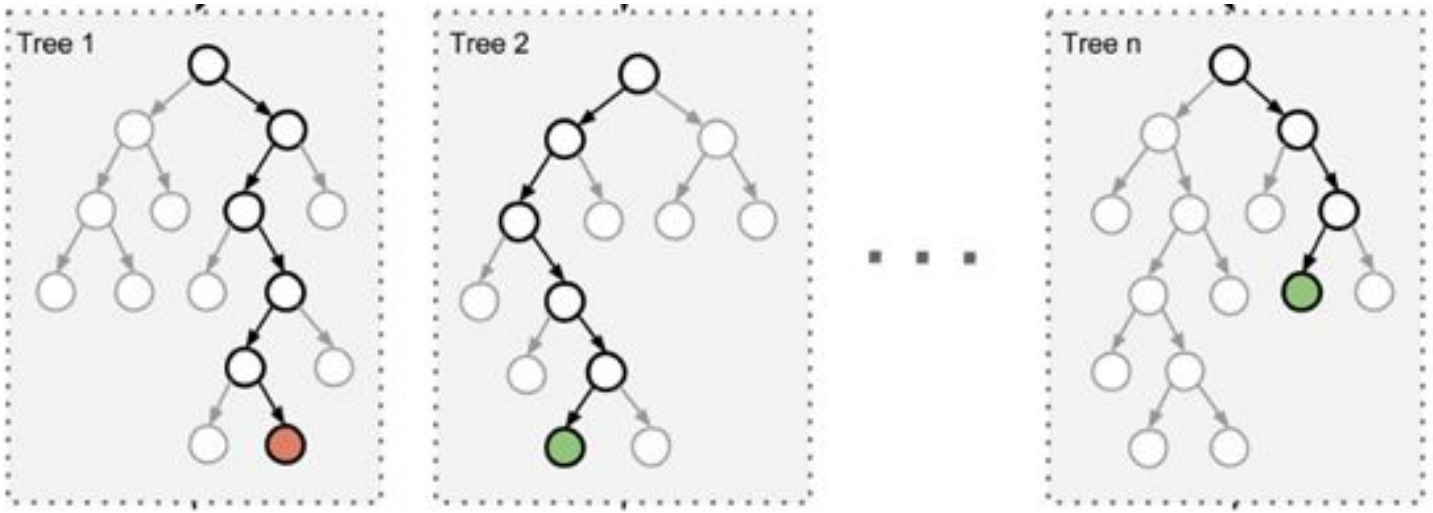
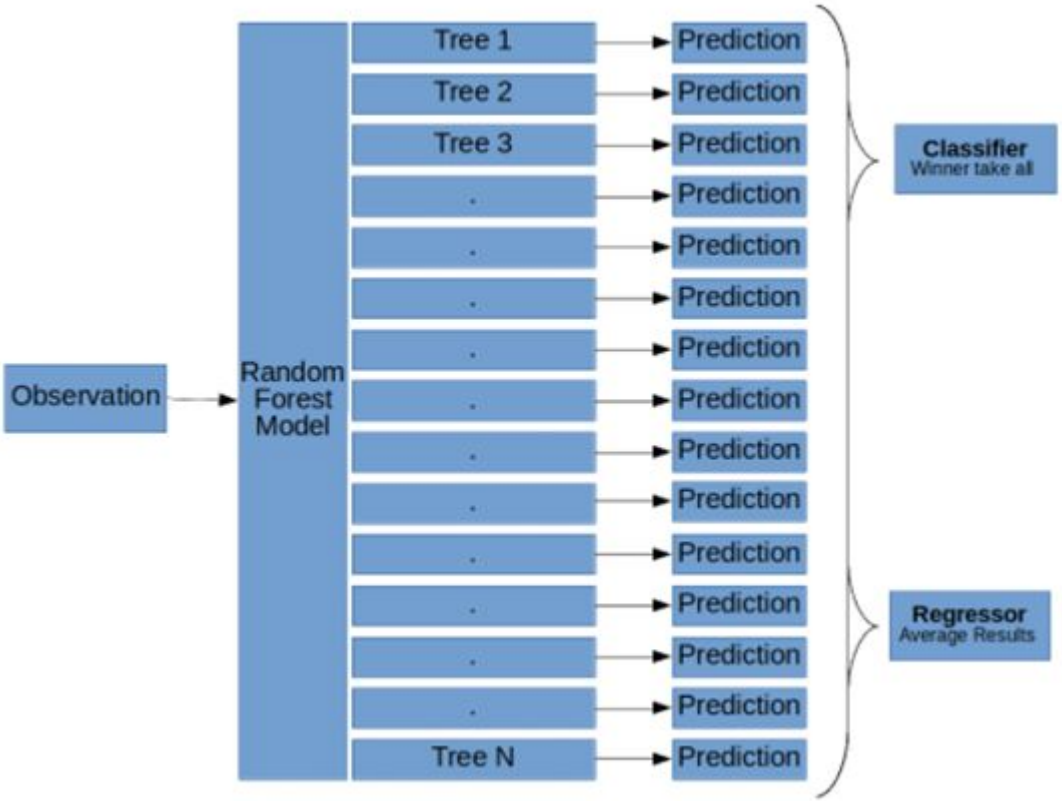


Image source: https://cdn-images-1.medium.com/max/1000/0*uGzCQfXIC-97VR10.



The bias of the forest increases slightly, but its variance decreases, an overall better model.

Image source: https://www.datascience.com/blog/random-forests-decision-trees-ensemble-methods

Random Forest vs Bagging

- Similarity :
  - Both use bootstrap sampling to obtain the data subsets for training the base learners

- Difference:
  - Random forest uses random subset of features to further randomize the tree
  - Bagging uses all the features

Classification Tree

```
from sklearn.tree import DecisionTreeClassifier

clf = DecisionTreeClassifier(criterion='gini',
max_depth=5, min_samples_leaf=60)

clf=clf.fit(Xtrain, encoded)
```

Difference between classification and regression tree model

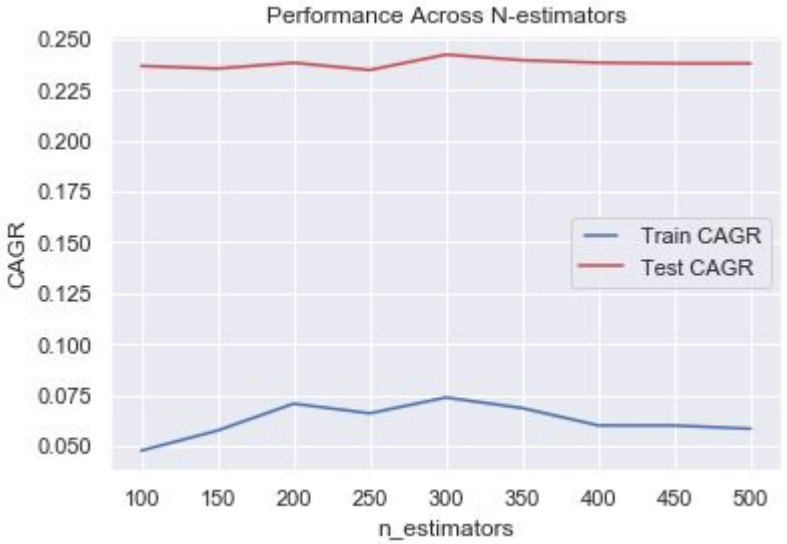
- Unordered value
- Ordered value
- Categorical
- Continuous

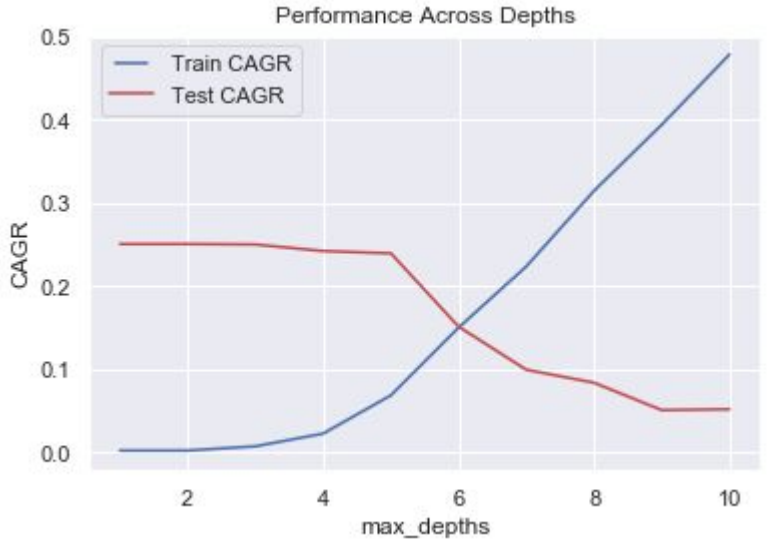
How to get results

```
from sklearn.metrics import r2_score

print(('R-squared = {}').format(r2_score(ytrain, Ypred)))
```

```
longs = pd.DataFrame(compare_nan_array(np.greater,
retPred, 0).astype(int)).shift(1)

shorts = pd.DataFrame(compare_nan_array(np.less,
retPred, 0).astype(int)).shift(1)
```

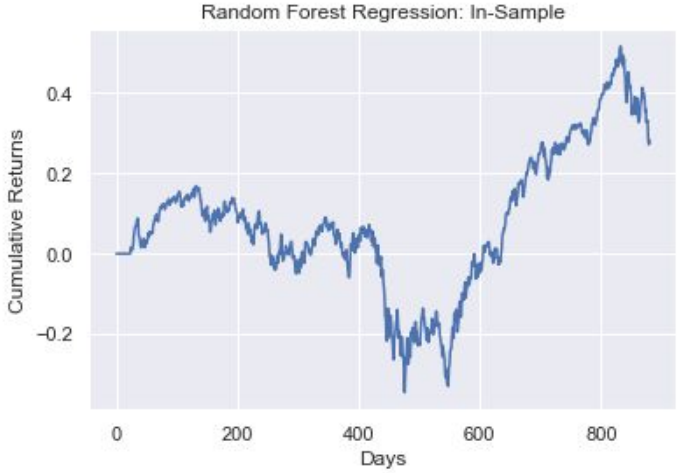
Results comparison-Random Forest





Which parameters to choose?

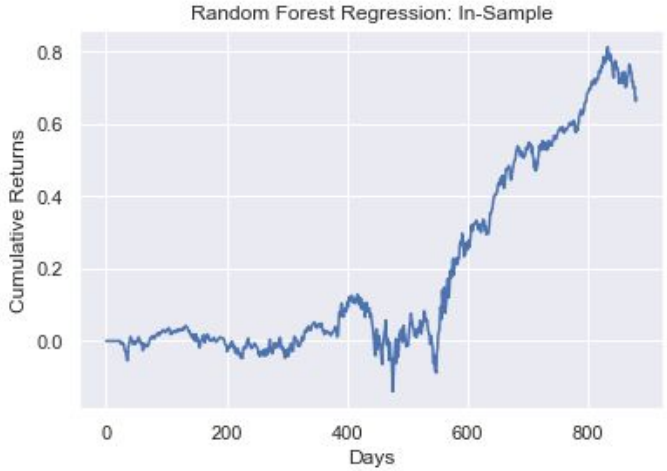
```
In-sample: CAGR=0.0738231 Sharpe ratio=0.397353 maxDD=-0.439698 maxDDD=531 Calmar ratio=0.167895
Out-of-sample: CAGR=0.242307 Sharpe ratio=1.37122 maxDD=-0.204809 maxDDD=171 Calmar ratio=1.18309
```



Random Forest

- The model is slightly underfitting
- Why?

```
In-sample: CAGR=0.15848 Sharpe ratio=0.847873 maxDD=-0.237785 maxDDD=210 Calmar ratio=0.666482
Out-of-sample: CAGR=0.15143 Sharpe ratio=1.31301 maxDD=-0.127081 maxDDD=118 Calmar ratio=1.1916
```



Random Forest

- Best fit
- Why?

Results comparison –Random Forest vs Classification Tree

Random Forest

- In-sample:
  - CAGR=0.15848 Sharpe ratio=0.847873 maxDD=-0.237785 maxDDD=210 Calmar ratio=0.666482
- Out-of-sample:
  - CAGR=0.15143 Sharpe ratio=1.31301 maxDD=-0.127081 maxDDD=118 Calmar ratio=1.1916

Classification Tree

- In-sample:
  - CAGR=-0.00842592 Sharpe ratio=0.108305 maxDD=-0.470765 maxDDD=687 Calmar ratio=-0.0178984
- Out-of-sample:
  - CAGR=0.222729 Sharpe ratio=1.25807 maxDD=-0.201063 maxDDD=171 Calmar ratio=1.10776

Fundamental Analysis vs Technical Analysis

Fundamental Analysis

Fundamental analysts study anything that can affect the security's value, including macroeconomic factors (e.g., economy and industry conditions) and microeconomic factors (e.g., financial conditions and company management).

Technical Analysis

Technical analysis is a trading discipline employed to evaluate investments and identify trading opportunities by analyzing statistical trends gathered from trading activity, such as price movement and volume.

Future Work

- Test our model on more datasets
- Improve our model by
  - More hyper-parameter tuning
  - Comparing with other models
    - Neural Network
    - Boosting Ensembles (sequential)
    - Gaussian Process
    - ...