

Training a Random Forest II

MGMT 638: Data-Driven Investments: Equity

Kerry Back, Rice University



Outline

- We follow 06b-random_forest_train to train the model but include mktvol interactions and set maxdepth=4.
- We save the trained model to forest_ver2.joblib.



Read data



```
In [1]: import pandas as pd

url = "https://www.dropbox.com/scl/fi/hjpebns5qv0nzh1uc14tr/data-2023-11-13.c
df = pd.read_csv(url)
df.head()
```

```
Out[1]:
```

	ticker	date	marketcap	pb	ret	mom	volume	volatility
0	AACC	2011-01-14	188.3	1.4	-0.014634	-0.184615	2.078000e+04	0.071498
1	AAI	2011-01-14	1012.1	2.0	0.002677	0.438224	2.775580e+06	0.128450
2	AAIC	2011-01-14	189.3	1.0	-0.010119	0.684547	3.466000e+04	0.048505
3	AAON	2011-01-14	479.4	4.2	0.007778	0.528685	2.817291e+05	0.044912
4	AATC	2011-01-14	63.3	1.4	-0.013960	0.008216	6.800000e+03	0.049756



Define model and target variable



```
In [2]: from sklearn.ensemble import RandomForestRegressor
forest = RandomForestRegressor(max_depth=4)

df["target"] = df.groupby("date", group_keys=False).ret.apply(
    lambda x: 100*x.rank(pct=True)
)
```

Define predictors (features)



```
In [3]: features = [  
        "marketcap",  
        "pb",  
        "mom",  
        "volume",  
        "volatility",  
        "roe",  
        "accruals",  
        "agr"  
    ]  
features.sort()
```



```
In [4]: for x in features:
        df[x+"_vol"] = df[x]*df.mktvol

        features += [x+"_vol" for x in features]
```

Filter to most recent 3 years



```
In [5]: dates = df.date.unique()
        dates.sort()
        df = df[df.date.isin(dates[-156:])]
```

Train the model



```
In [6]: forest.fit(X=df[features], y=df.target)
```

```
Out[6]: ▼ RandomForestRegressor
```

```
RandomForestRegressor(max_depth=4)
```



Save the model



```
In [7]: # Local machine version  
  
from joblib import dump  
dump(forest, "forest_ver2.joblib")
```

```
Out[7]: ['forest_ver2.joblib']
```

In []: *# colab version*

```
from joblib import dump
from google.colab import drive
drive.mount('/content/drive')
dump(forest, "/content/drive/My Drive/forest_ver2.joblib")
```

