FITTING RANDOM FORESTS

RANDOM FOREST

RANDOM FOREST

Random Forests are

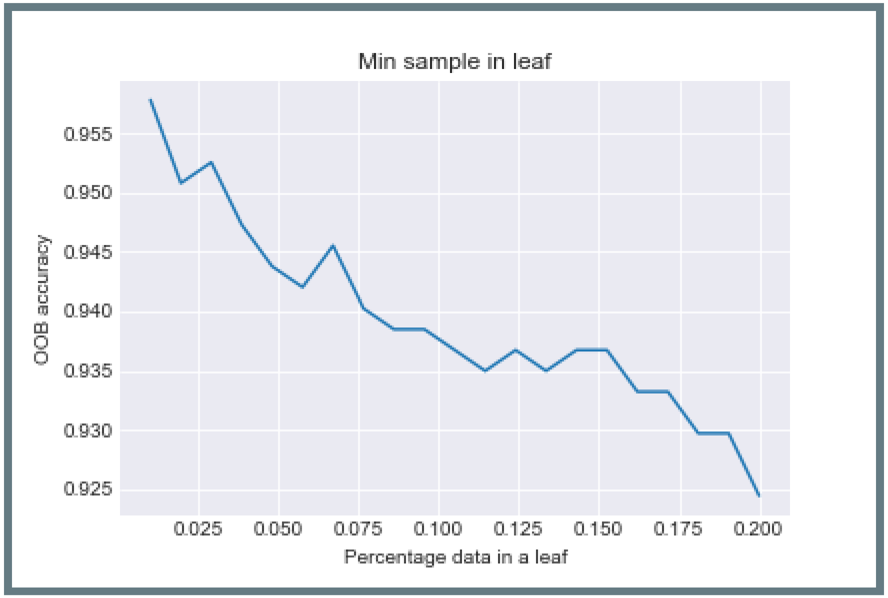
- bagged decision trees where
- a random subset of predictors is used for each split

RANDOM FORESTS

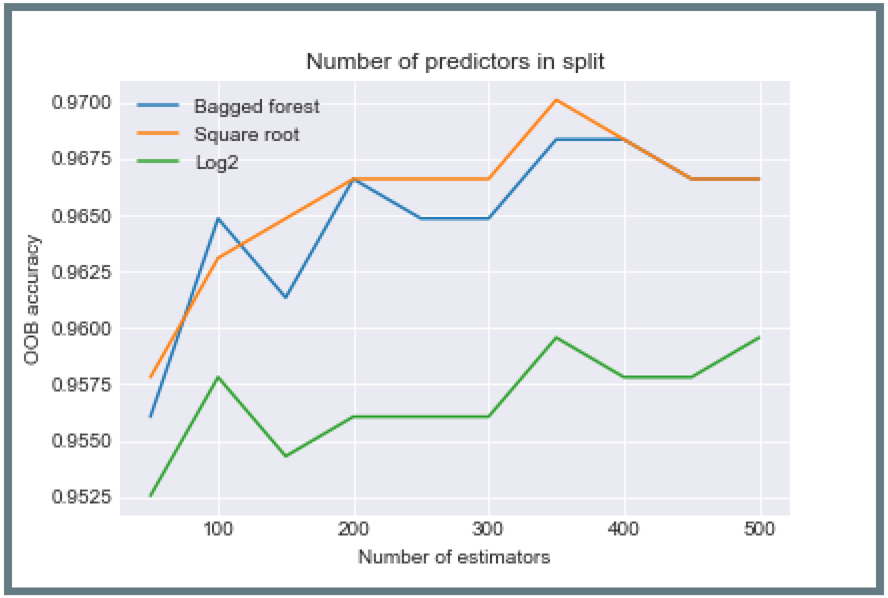
There are four main parameters for a random forest:

1. Number of estimators (bootstrap samples)
2. Proportion of predictors to use at each split (`max_features`)
3. Depth of each tree (`max_depth`)
4. Minimum sample size at each terminal node (`min_sample_leaf`)

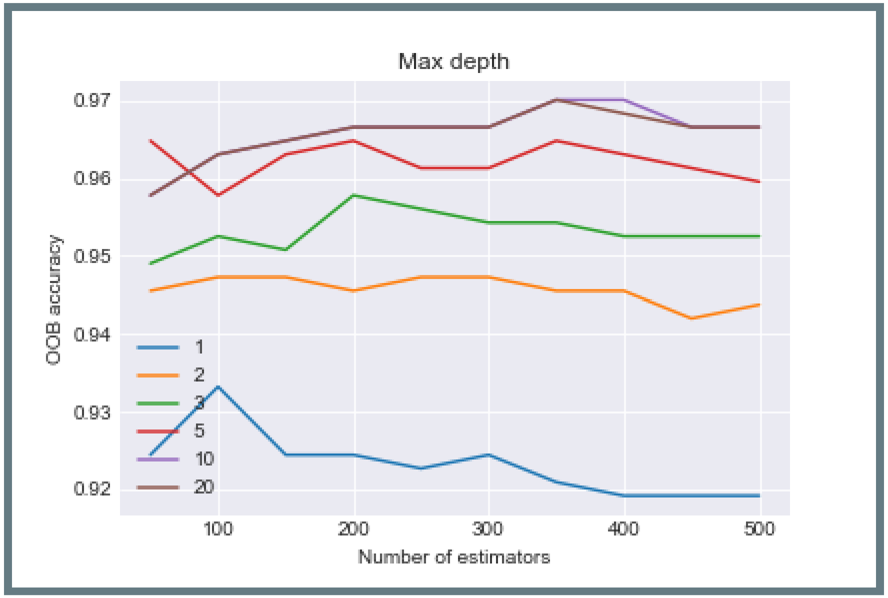
MIN_SAMPLE_LEAF



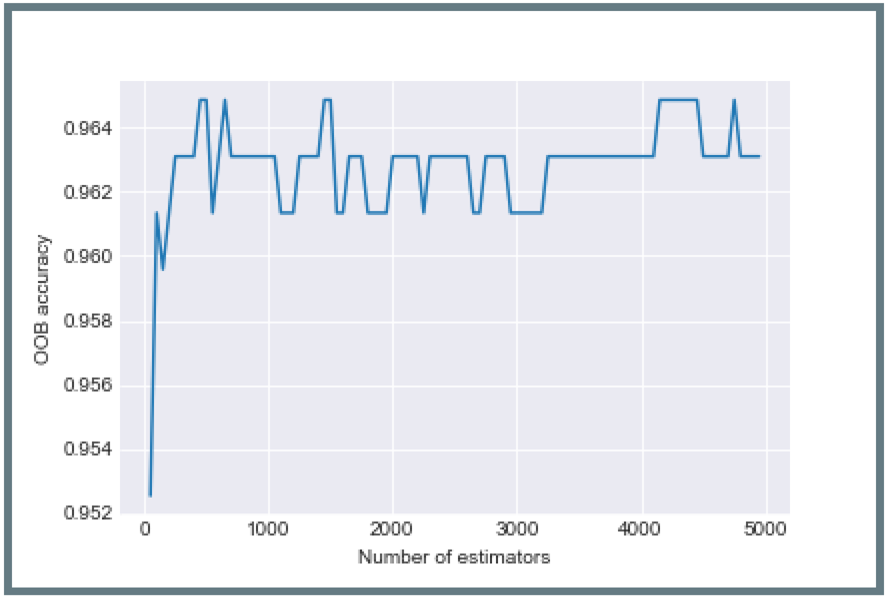
MAX_FEATURES



MAX_DEPTH


Max depth
0.97
0.96
0.95
0.94
0.93
0.92
OOB accuracy
100
200
300
400
500
Number of estimators
1
2
3
5
10
20

NUMBER OF ESTIMATORS



FEATURE IMPORTANCE AND SELECTION

FEATURE IMPORTANCE

By default, RandomForestClassifier gives the improvement in Gini Index due to splitting on a predictor as the feature importance

- It is influenced by the number of categories in each variable

PERMUTATION IMPORTANCE

The permutation importance is better, but not directly implemented

See implementation in script

FEATURE IMPORTANCE IN GENERAL

- One at a time
- Can be influenced by collinearity/dependence
- Features which are collinear will tend to depress each others’ importance score
- Can be influenced by granularity of the predictor
- Features with more levels are more likely to be “important”

VARIABLE SELECTION

- Use cross-validation and backward selection to optimize model

TAKING ADVANTAGE OF BAGGING

AN EXPLORATION OF DIFFERENT METRICS

1. Accuracy : What proportion of predictions are correct
2. Precision: What proportion of positives are true
3. Recall: What proportion of true positives are called positive
4. AUC : Area under the receiver operating characteristic (ROC) curve

- ROC maps (1-specificity) against sensitivity for different cutoffs
- Sensitivity = Recall
- Specificity = What proportion of true negatives are called negative

   5. F1 score: harmonic mean of Precision and Recall

   6. Brier score: Mean squared error of probability

predictions

PROBABILITY PREDICTIONS

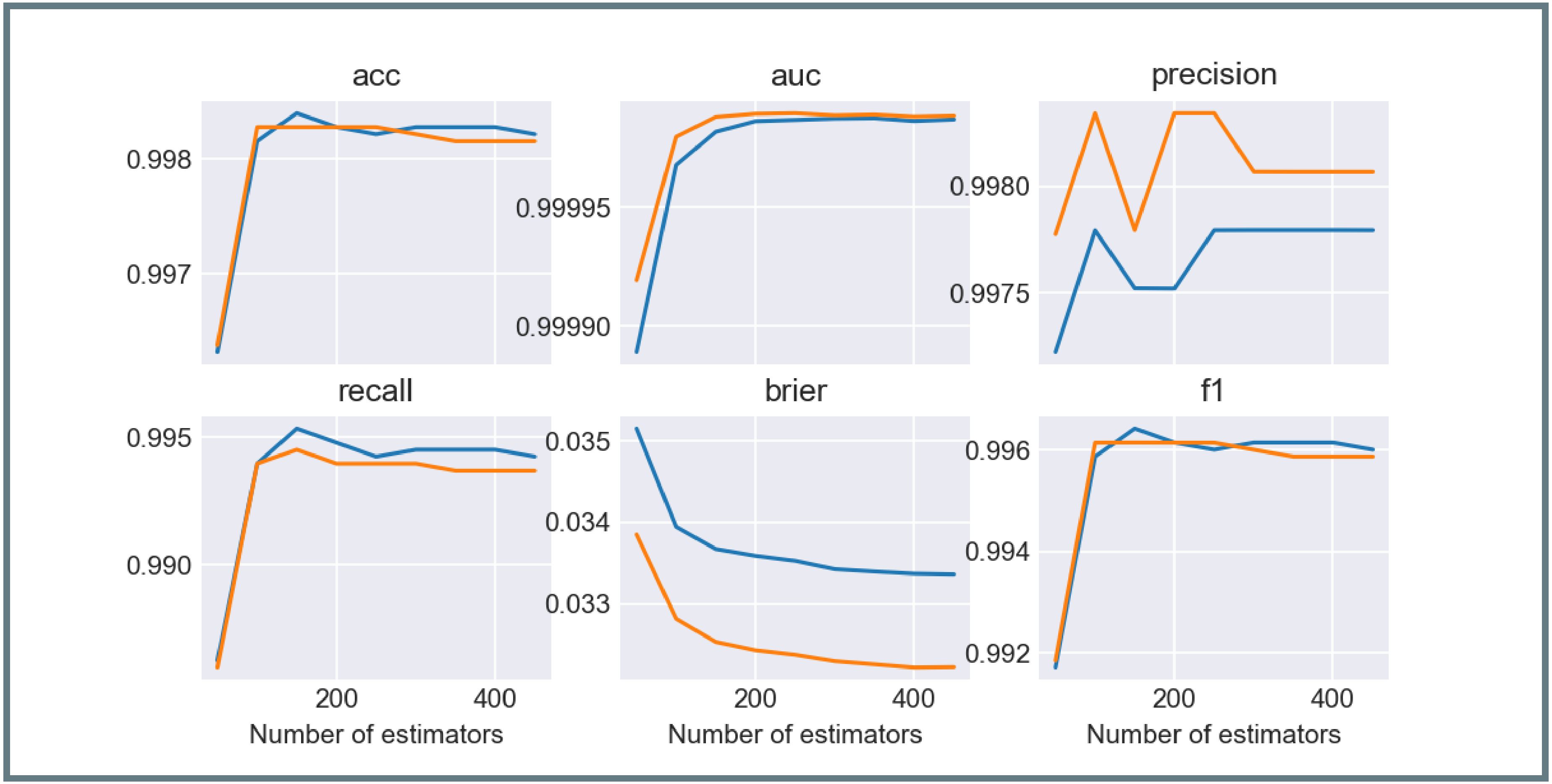
You can get probability predictions from random forest classifiers

Proportion of ‘1’ calls for each point

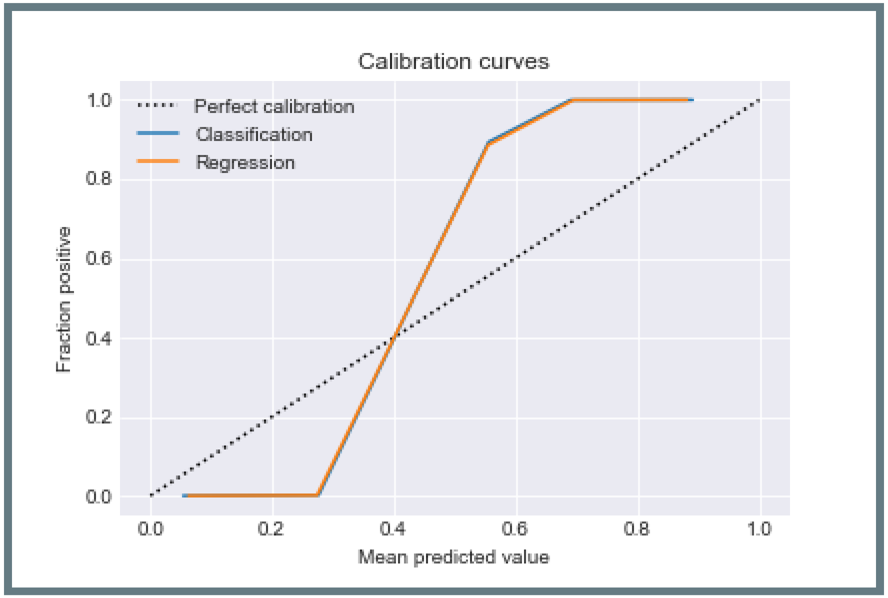
PROBABILITY PREDICTIONS

For binary outcomes, you can also get probability predictions from random forest regressors

PERFORMANCE



CALIBRATION CURVES



STACKING