
H2O.ai

Ensembles

What it is:

- "Ensemble methods use multiple learning algorithms to obtain better predictive performance that could be obtained from any of the constituent learning algorithms." (Wikipedia, 2015)
- Random Forests and Gradient Boosting Machines (GBM) are both ensembles of decision trees.
- Stacking, or Super Learning, is technique for combining various learners into a single, powerful learner using a second-level metalearning algorithm.

What it's not:

Ensembles typically achieve superior model performance over singular methods. However, this comes at a price — computation time.

Case Study: Lending Club Dataset

- Loan data from 2007 up until 2015 including rejected applications and accepted applications.
- Of the 500k accepted applicants about 160k loans have either been completely paid off or defaulted.
- There are about 4 million applicants in the rejected loans dataset.
- **Use Case 1**: Predict the likelihood of a user defaulting based on the information supplied when applying for a loan.
- **Use Case 2**: Determine the interest rate Lending Club would have offered the user based on the information supplied when applying for a loan.
- Full Data: https://www.kaggle.com/wendykan/lending-club-loan-data
- H2O Subset: https://s3.amazonaws.com/h2o-public-test-data/bigdata/laptop/lending-club/loan.csv

Ensembles

What it is:

- "Ensemble methods use multiple learning algorithms to obtain better predictive performance that could be obtained from any of the constituent learning algorithms." (Wikipedia, 2015)
- Random Forests and Gradient Boosting Machines (GBM) are both ensembles of decision trees.
- Stacking, or Super Learning, is technique for combining various learners into a single, powerful learner using a second-level metalearning algorithm.

What it's not:

Ensembles typically achieve superior model performance over singular methods. However, this comes at a price — computation time.