# Random Forests

Session 8 In-Class Slides

# Intended Learning Outcomes?

Identify aspects of business problems that cause standard analytics models to become useless or less effective.

Apply advanced techniques to overcome or mitigate the weaknesses of standard analytics models.

Evaluate performance of the advanced predictive techniques.

Explain the workings and results of the advanced predictive techniques in the context of the business problem to client/employer.

Propose business solutions/recommendations based on the advanced predictive techniques.

# Quiz

Ungraded. Check your understanding of this Session Content.
Use your real name (not nickname) in the quiz.

# Discuss Solution to Exercise 1

Pre-class Activity

# Do Exercise 2

In-class

# Python RF Implementation

- Categorical Y:
  - sklearn.ensemble.RandomForestClassifier
  - https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- Compared to R randomForest():
  - Python used default B = 100. Please override to 500.
  - Python used same default RSF size = sqrt(num of X variables).
  - Used Mean Gini Increase as measure of variable importance. Permutation based approach not avail within this library. Need to import another python library (e.g. rfpimp).

# Python RF Implementation

- Continuous Y:
  - sklearn.ensemble.RandomForestRegressor
  - https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
- Compared to R randomForest():
  - Python used default B = 100. Please override to 500.
  - Python used default RSF size = num of X variables. Please change to int(num of X variables/3).
  - Permutation based approach for assessing variable importance not avail within this library. Need to import another python library (e.g. rfpimp).

# Summary

- Random Forest
  - Bagging
    - Bootstrap samples
    - One maximal CART per Bootstrap sample.
  - Random Subset Feature Selection
    - Default size int(M/3) for continuous Y
    - Default size int(sqrt(M)) for categorical Y
  - Errors calculated from OOB cases
    - Check errors stabilised before reaching ntree (default 500).
  - Permutation based approach for assessing variable importance

# What did we learn in this topic?

Identify aspects of business problems that cause standard analytics models to become useless or less effective.

Apply advanced techniques to overcome or mitigate the weaknesses of standard analytics models.

Evaluate performance of the advanced predictive techniques.

Explain the workings and results of the advanced predictive techniques in the context of the business problem to client/employer.

Propose business solutions/recommendations based on the advanced predictive techniques.

# Reminder

Please complete the Pre-Class Learning Activities before next class.

# Reflection on your Learning

| | |
|---|---|
| Go | NTULearn Class Site > Journal |
| Post | Read the instructions and post entry on this week's learning.<br>• Reply on the 3 questions as stated in the Journal Instructions. |