University of Essex

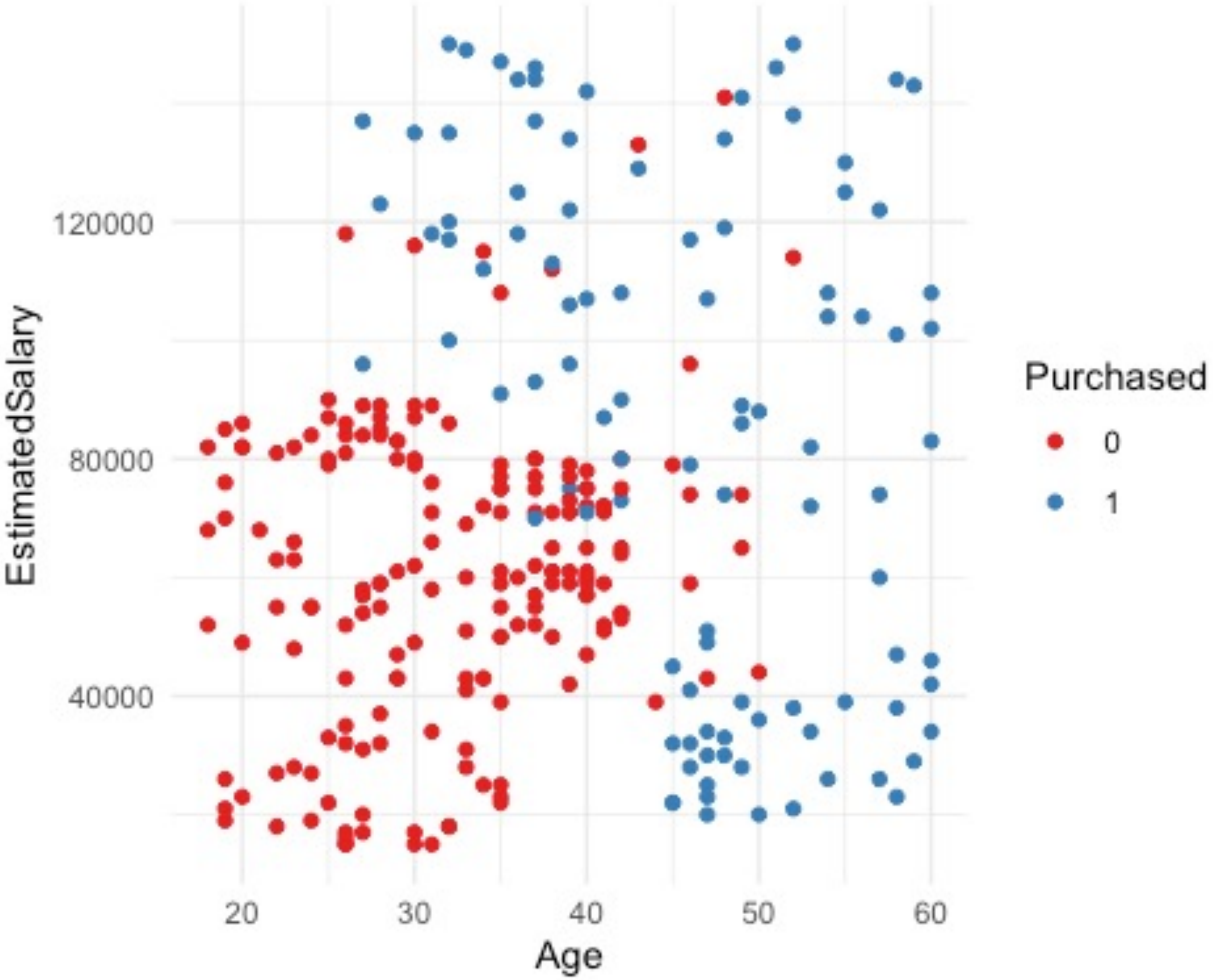
Tree Based Methods

Akitaka Matsuo

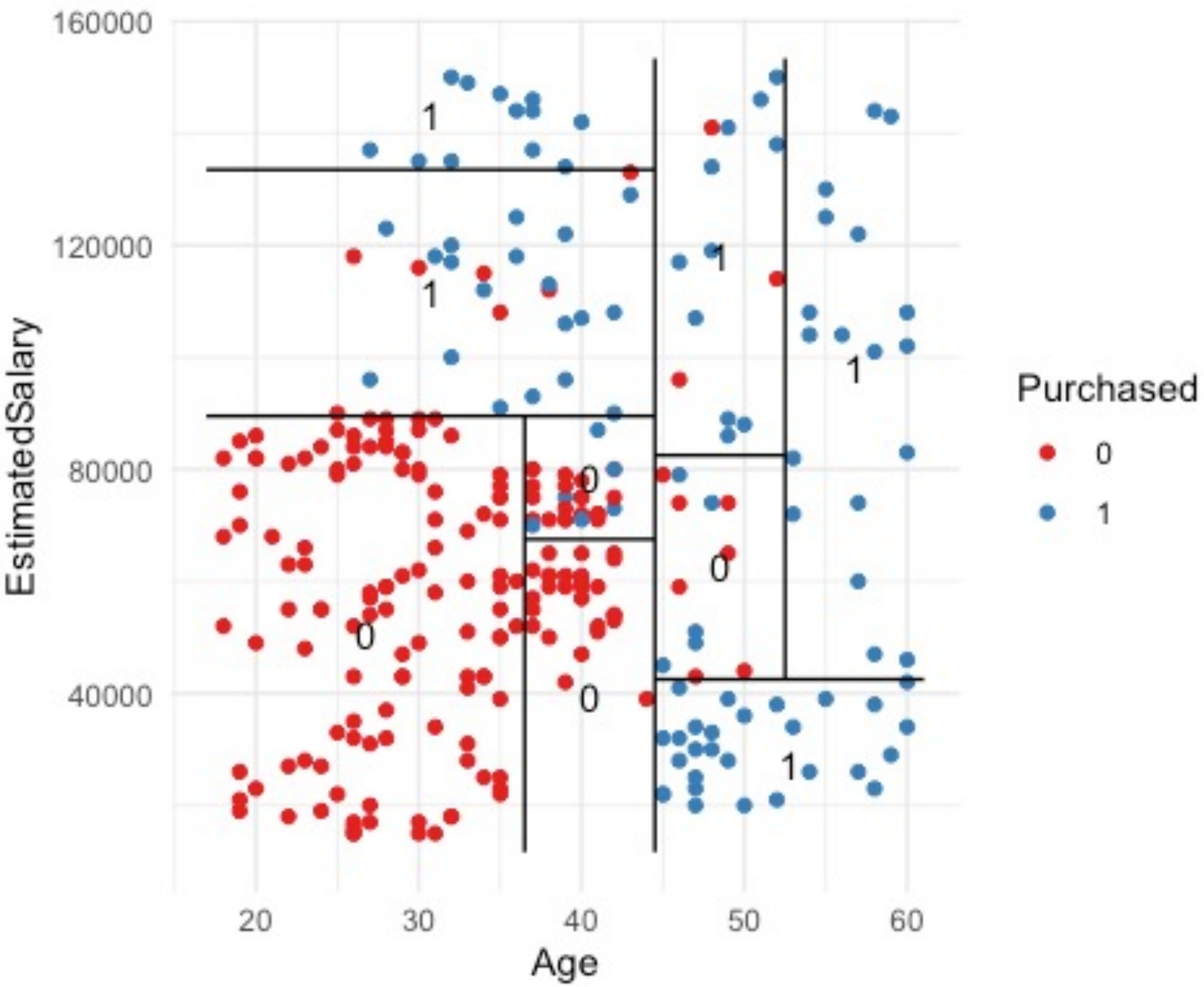
Content

- – Tree based models
    - Idea
    - Simple tree
    - Random Forest
    - AdaBoost
- – Final Notes

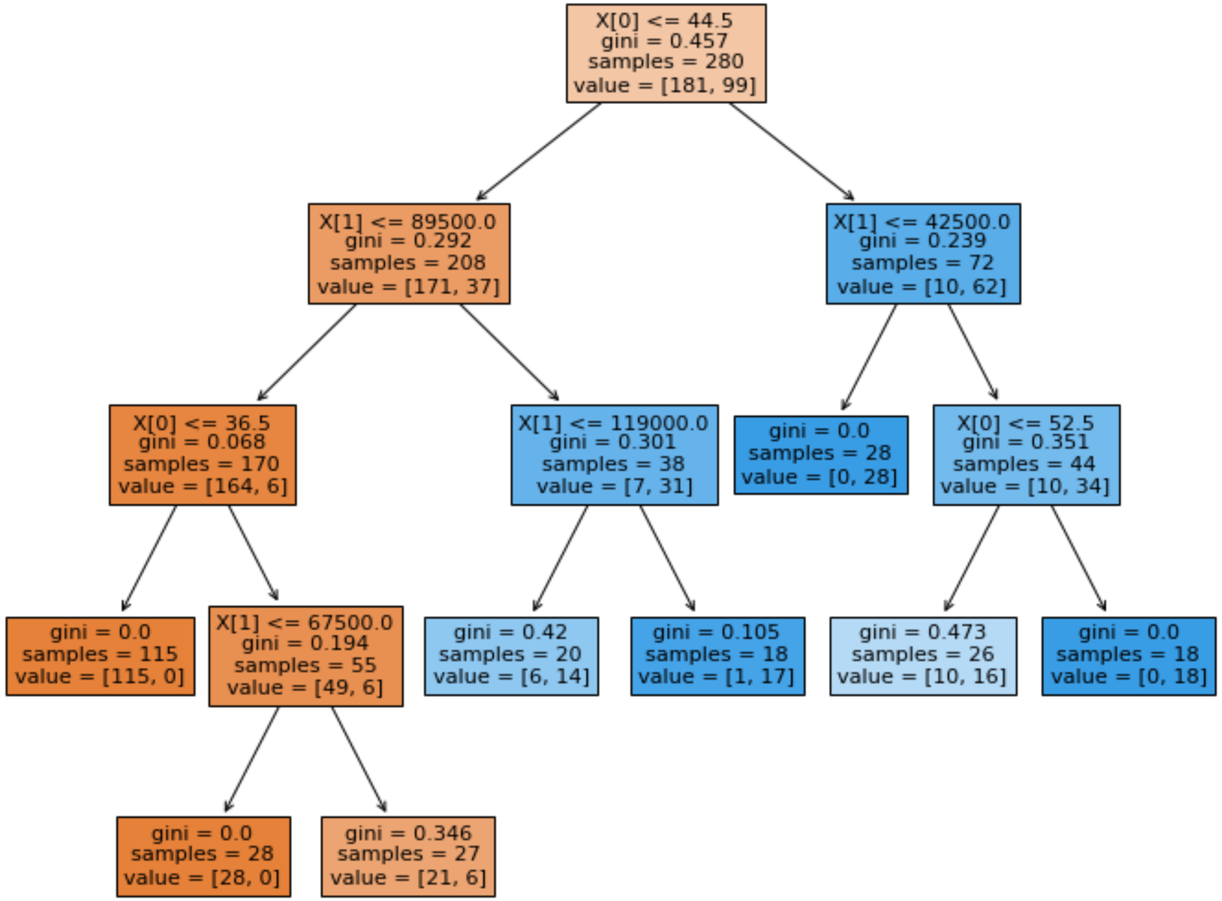
Tree based method: example



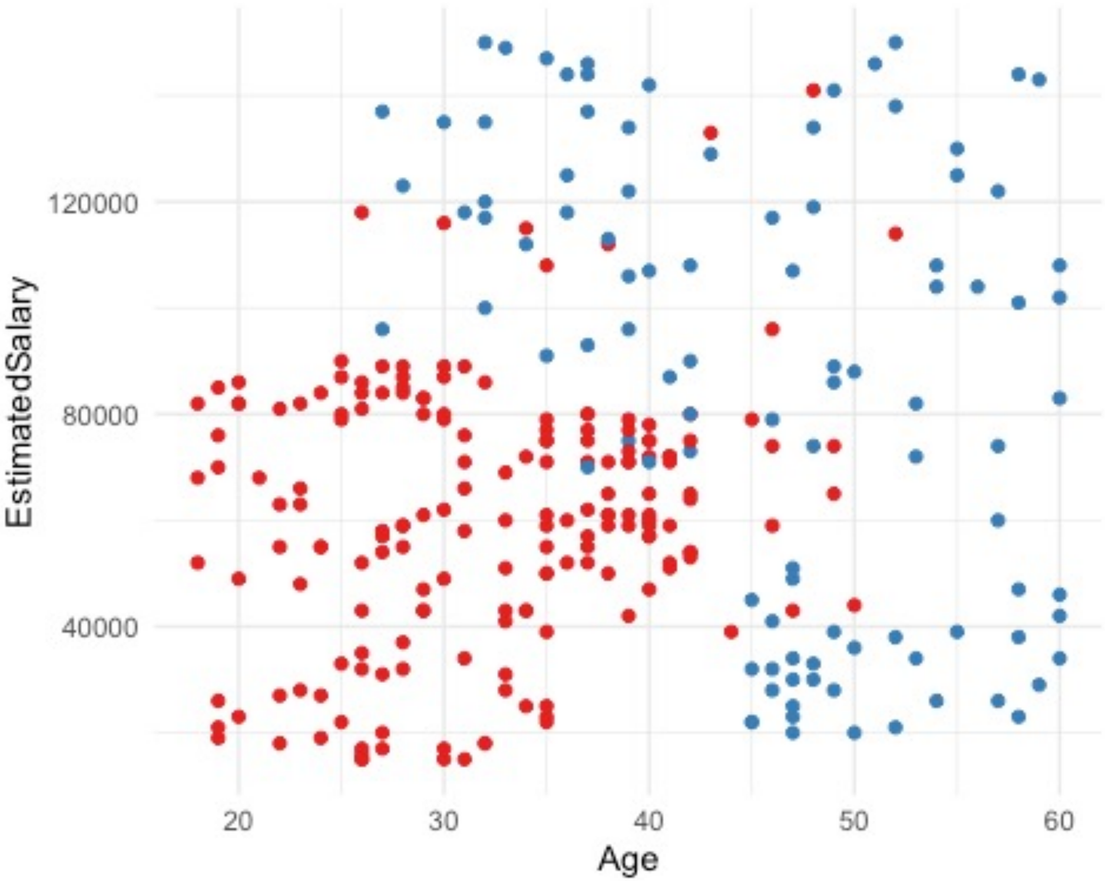
Tree output: Decision boundary



Tree output: Decision tree

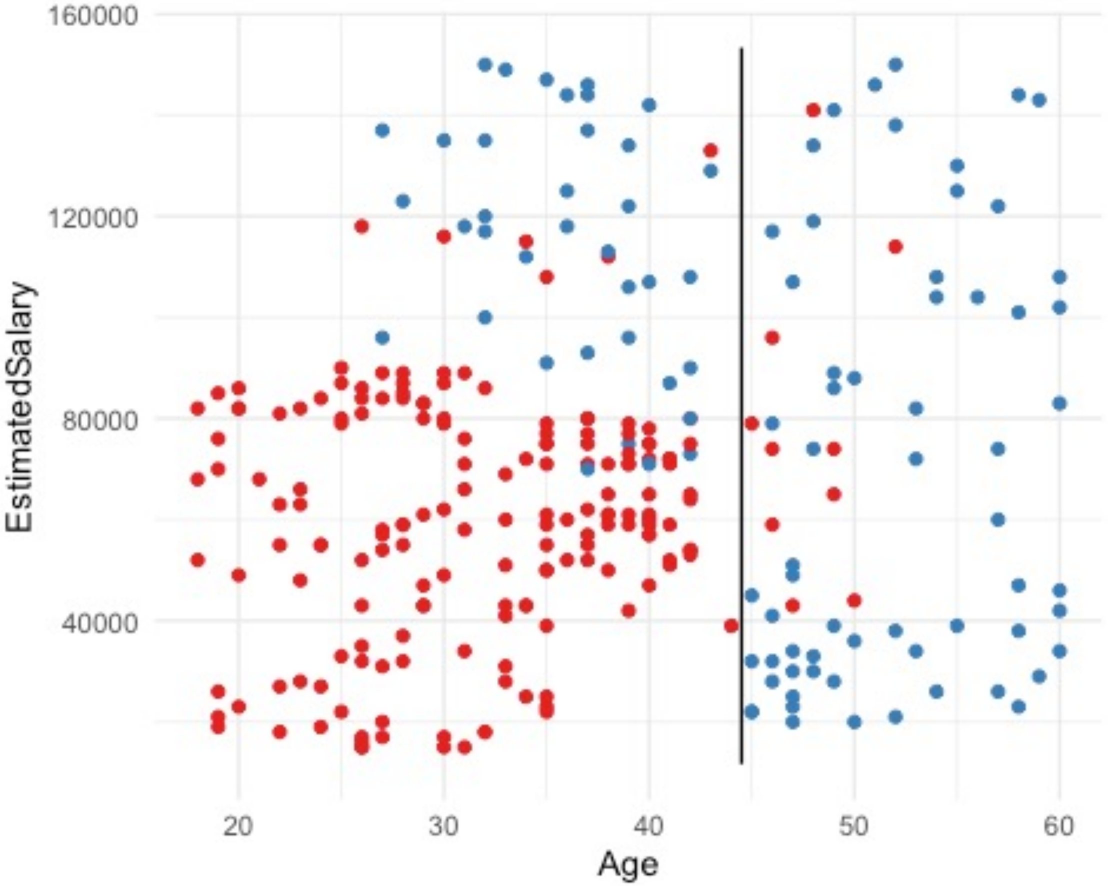


Tree making process



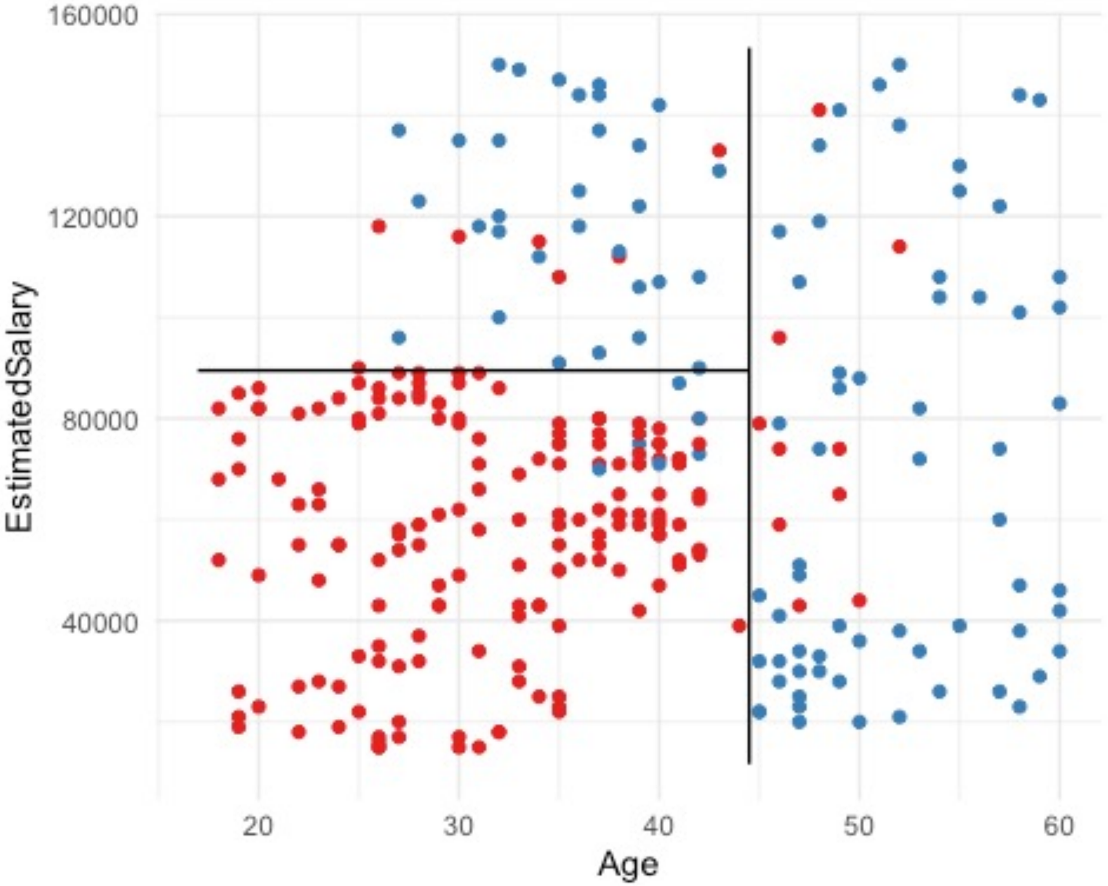
- Looking at the figure, think about where to put a vertical or horisontal line that separate two classes the best

Tree making process (1)



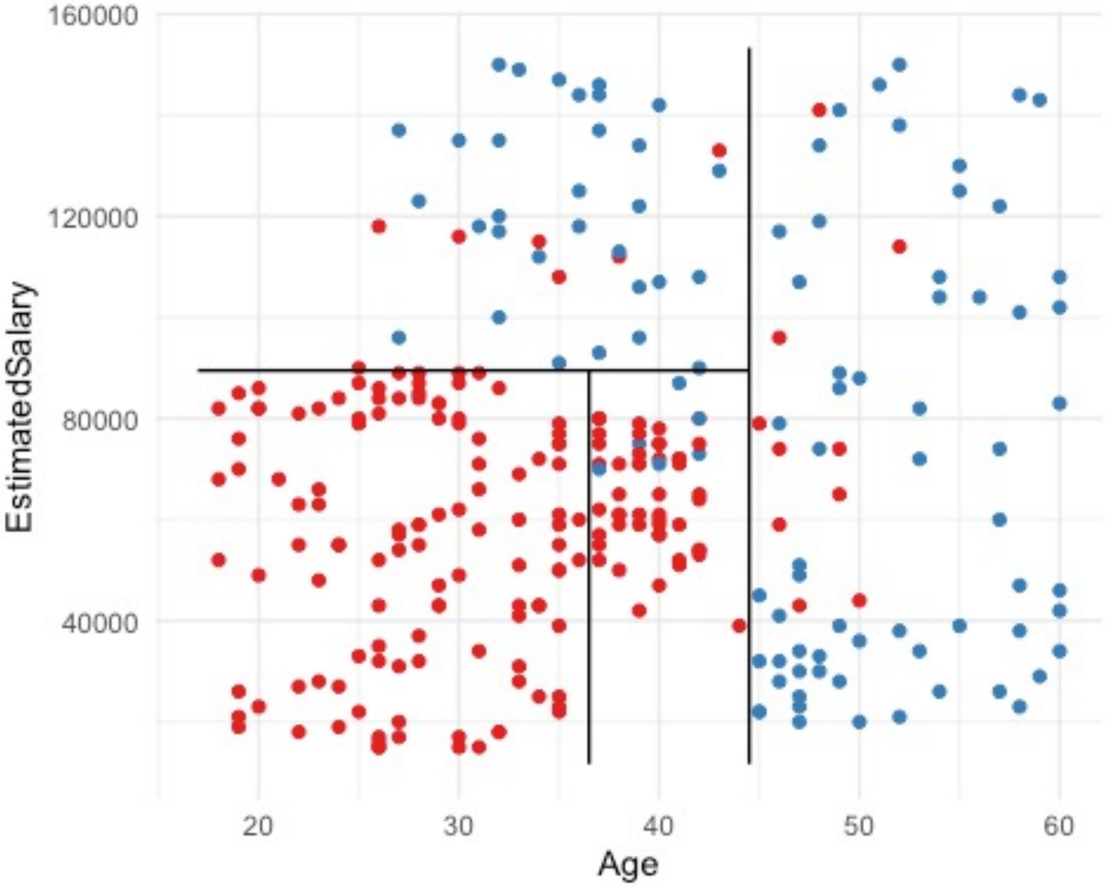
- A vertical line at the age of 45 split the space well
- Look at the left space, where to put a vertical or horizontal line?

Tree making process (2)



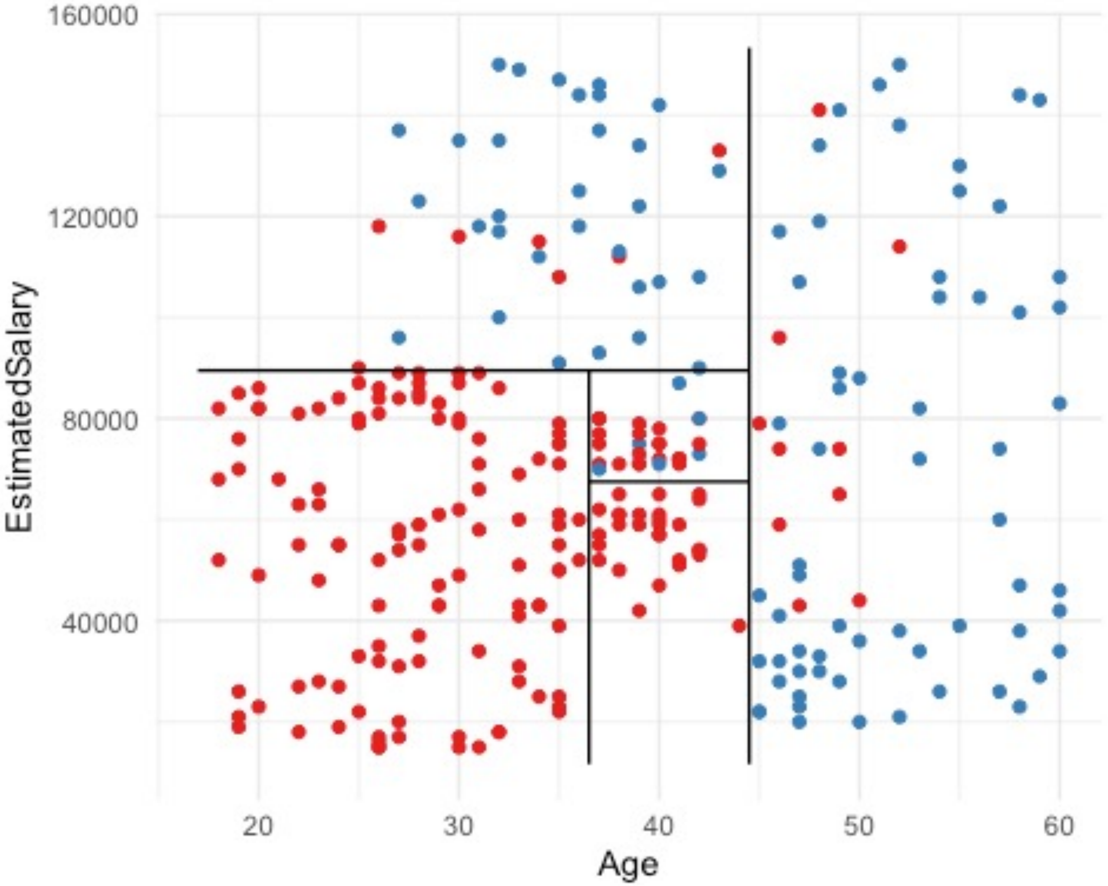
- A horizontal line with ~90K salary
- Left bottom space has still a lot of points
- Any way to split

Tree making process (3)



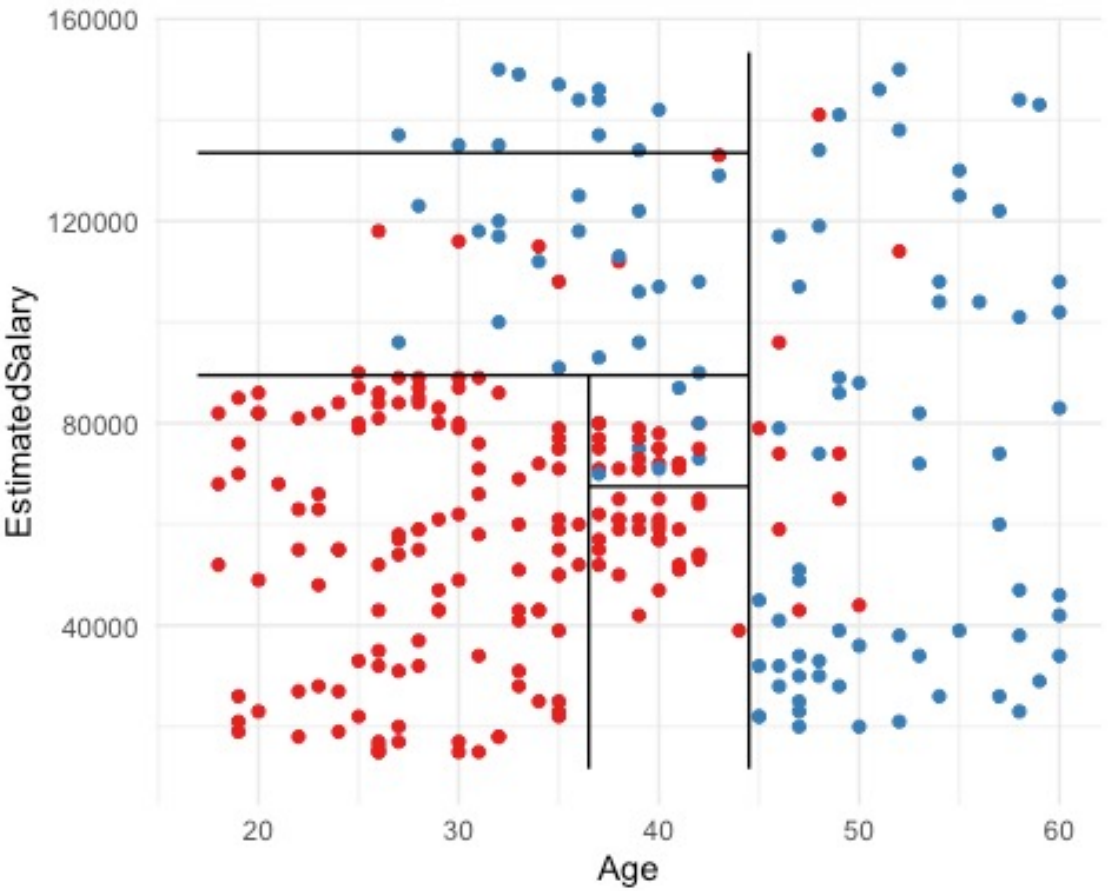
- Now the points in the bottom left are all red, so no further split there
- The bottom middle, any split?

Tree making process (4)



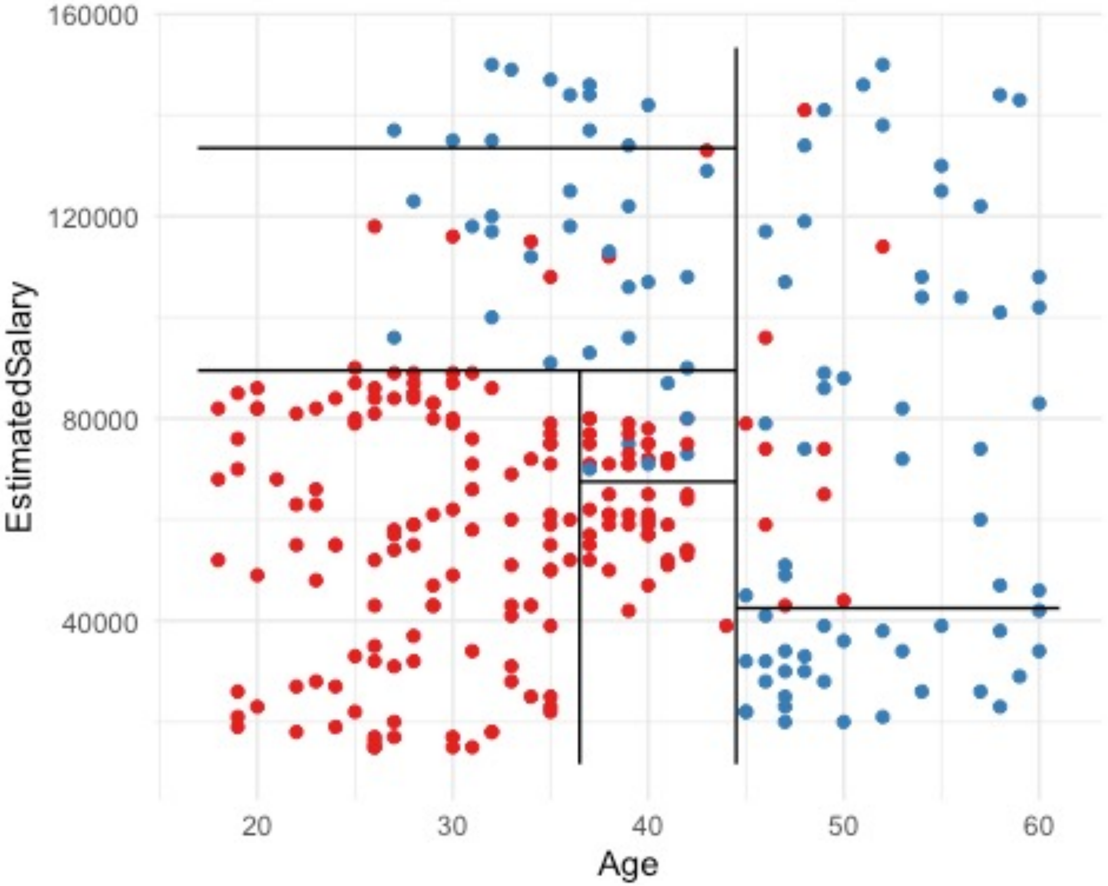
– It was actually possible, but there seems no further improvement is possible for the bottom left (three) regions

Tree making process (5)



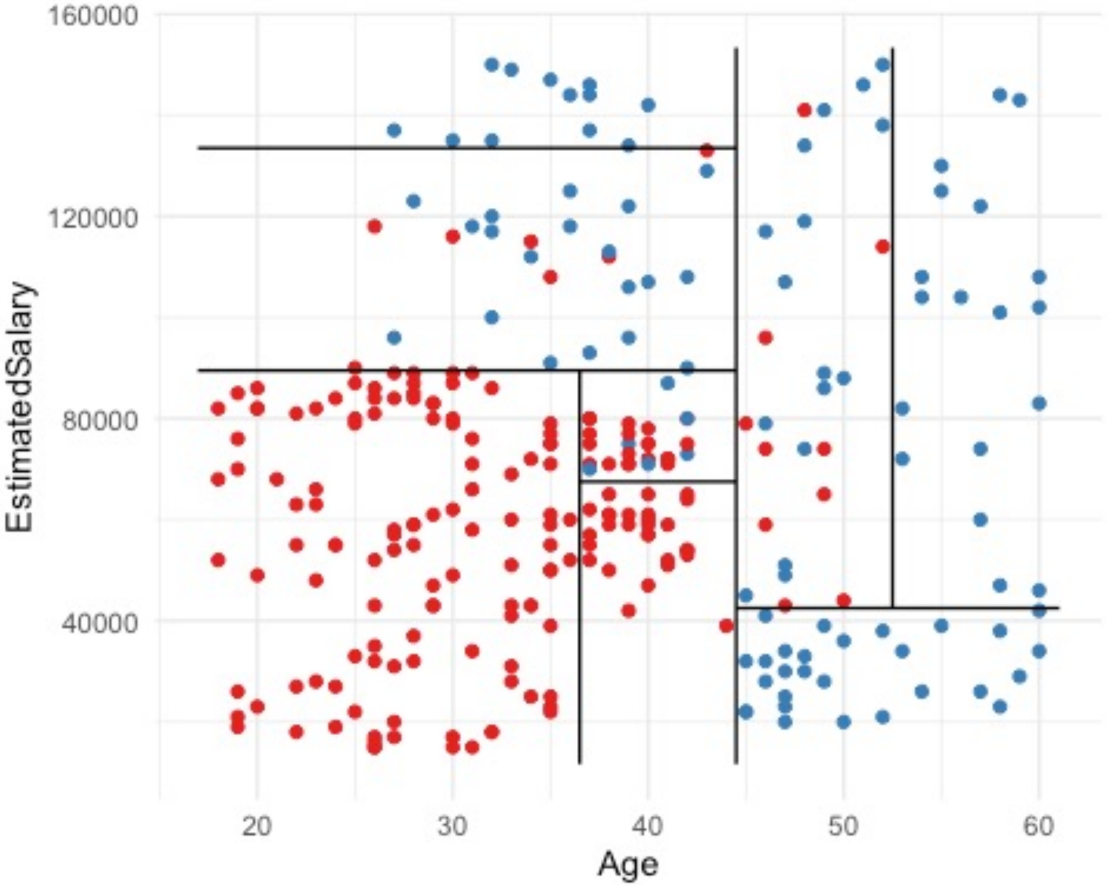
- Now top left is divided, but not much improvement
- Let’s think about the right half

Tree making process (6)



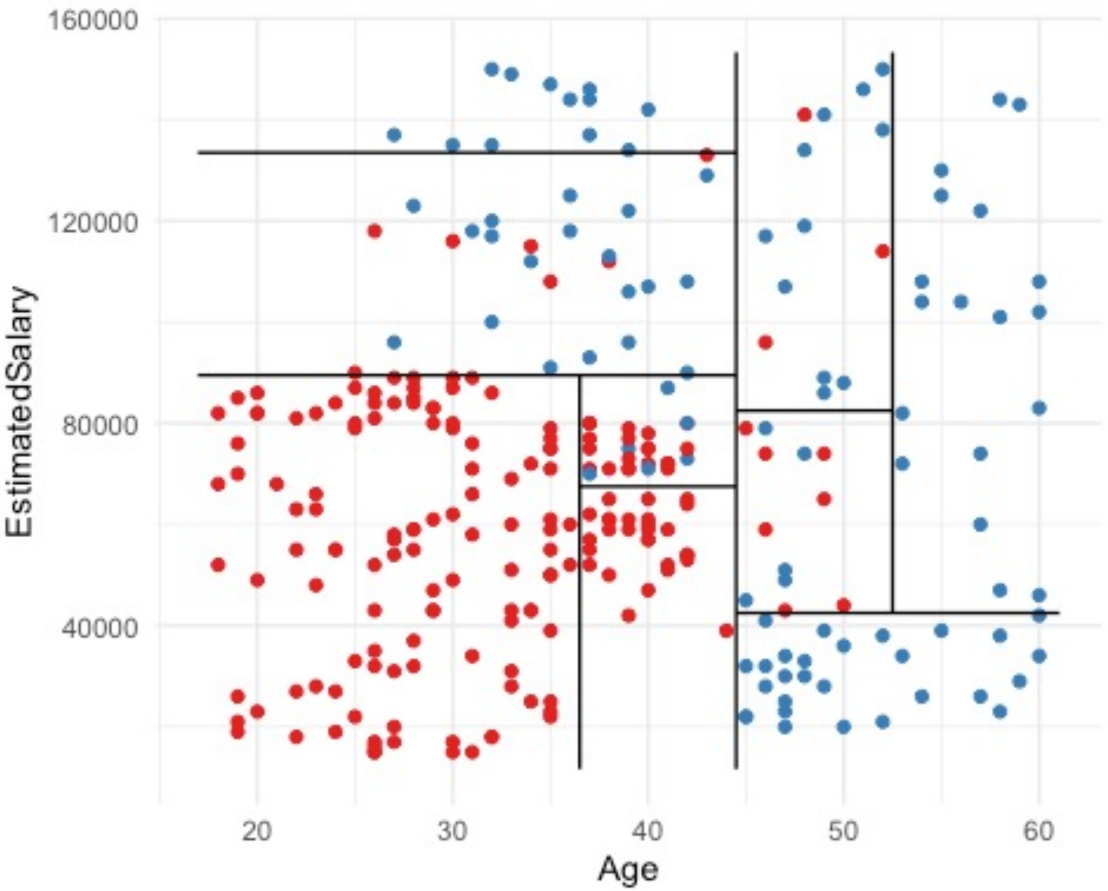
– Keep going

Tree making process (7)



- Keep going

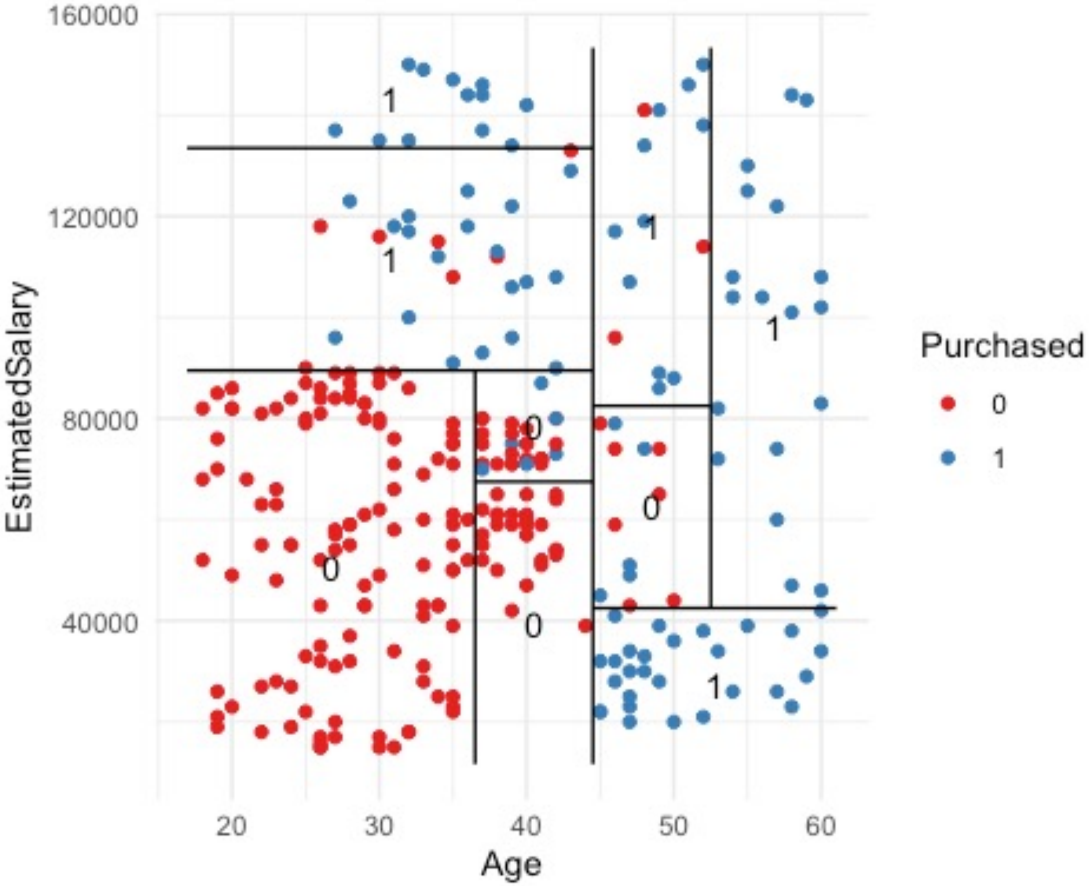
Tree making process (8)



– This is the point where no improvement is possible

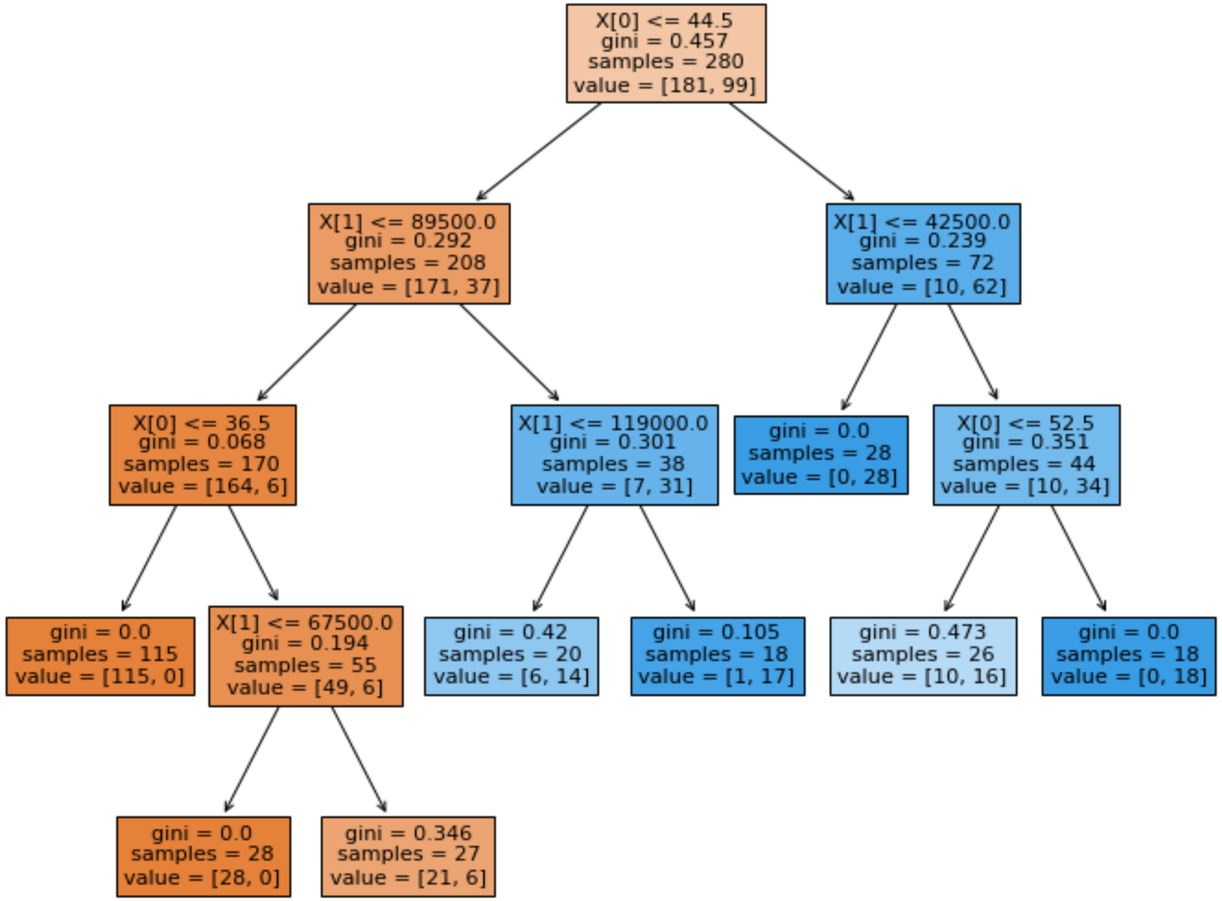
– So, make the prediction for each region

Final tree



– Tree is complete…

Tree output: Decision tree



Decision tree building

1. For a region, check each input variable for the best cutoff point
   - “Best” means the largest reduction of average impurity
     - – Gini: $1 - (\frac{N_1}{N})^2 - (\frac{N_2}{N})^2$
   - For an binary input variable (i.e. Gender), the cutoff is already set
2. Select the best variable among **all** variables
   - Again, based on the impurity
3. Split the space at the cutoff of the best variable
4. Do 1.-3. for sub-spaces
5. Stop splitting when no improvement is possible (e.g. Only one class exist in the region)

Decision tree classification

Pros

- so intuitive
- easy to calculate (for a computer)

Cons

- Test-set accuracy tends to be lower (the model tend to have high variance)
    - especially for a model with lots of inputs

Improvement

- So we need some sort of improvement to the model
- There are many methods, but intuitively, the improvement is done by estimating a lot of trees with different shapes, and aggregating the predictions of these trees
- **Methods**
  - Random forest
  - AdaBoost
  - Gradient boost
  - XGboost

Bagging

- – Build are lot of trees
- – How?
  - Bootstrapping the data
  - Build a tree
- – After making a lot of trees, predict each observation based on the majority rule

Random forest

- Build are lot of trees, similar to Bagging
- How?
  - Bootstrapping the data
  - Build a tree, but only consider small subset of variables each time
    - m randomly selected variable at each node (Typically $m = \sqrt{p}$, where p = num of input variables)
    - to get wide variety of trees (decorrelating)
- After making a lot of trees, predict each observation based on the majority rule
- Tuning parameter:
  - `mtry`: number of variables consider each time

AdaBoost

- Estimate a number of tiny trees sequentially
  1. Estimate a shallow tree with observation weights (1-3 levels)
  2. Evaluate the shallow tree (accuracy of prediction)
  3. Recalculate the weights of observations
     - increase the weight of the misclassified
     - decrease the weight of the correctly classified
  4. repeat 1.-3.
- Tuning parameters:
  - Number of iterations
  - Depth of tree
  - Learning rate

Final Note

- For the use of tree based methods see:
  - Montgomery, Jacob M, and Santiago Olivella. 2018. “Tree-Based Models for Political Science Data.” *American Journal of Political Science* 62(3): 729–44.
    - In depth review of tree based methods and application to the big data study
- For AdaBoost
  - https://www.youtube.com/watch?v=LsK-xG1cLYA
- For random forest
  - https://www.youtube.com/watch?v=J4Wdy0Wc_xQ