Random Forests in Actuarial Practice

First Part


Suguru Fujita（AKUR8）


Yasuhiro Matsue（PwC）

Speakers

- **Suguru Fujita**, FIAJ, ASA, CSPA, CERA

- **Yasuhiro Matsue**, AIAJ

AKUR8

公益社団法人日本アクチュアリー会
Think the Future, Manage the Risk

pwc

公益社団法人日本アクチュアリー会
Think the Future, Manage the Risk

Disclaimer: *The views expressed in this presentation are solely of the speakers and/or authors and are NOT affiliated with any organization.*

Agenda

1. Random Forests Basics (Suguru)

2. Properties Required for Actuarial Analysis
   Comparison with Other Methods
   Potential Use & Issues of RF in Actuarial Practice (Yasu)

Notation

Notation	Description
$\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$	Training data; each sample $Z_i \coloneqq (X_i, Y_i)$ is assumed to be i.i.d.
X_i	p-dimensional feature vector
Y_i	Response variable
(x, y)	New observation
n	Sample size
p	Number of features
B	Number of trees comprising Random Forests

Random Forests Basics

What's Random Forests?

Random Forests are:

- A method of combining decision trees with bagging method, a type of ensemble learning;
- A method that could be used for both regression and classification problem.

Key Concepts

1. Forests

- A bunch of decision trees

2. Randomness

- Bagging (Bootstrap AGGregatING): each tree is built from a different bootstrap sample of the training data.
- Features used for branching: in each tree, a certain number of features are randomly selected for each branching and the best one is used for the branching.

Random Forests Algorithm

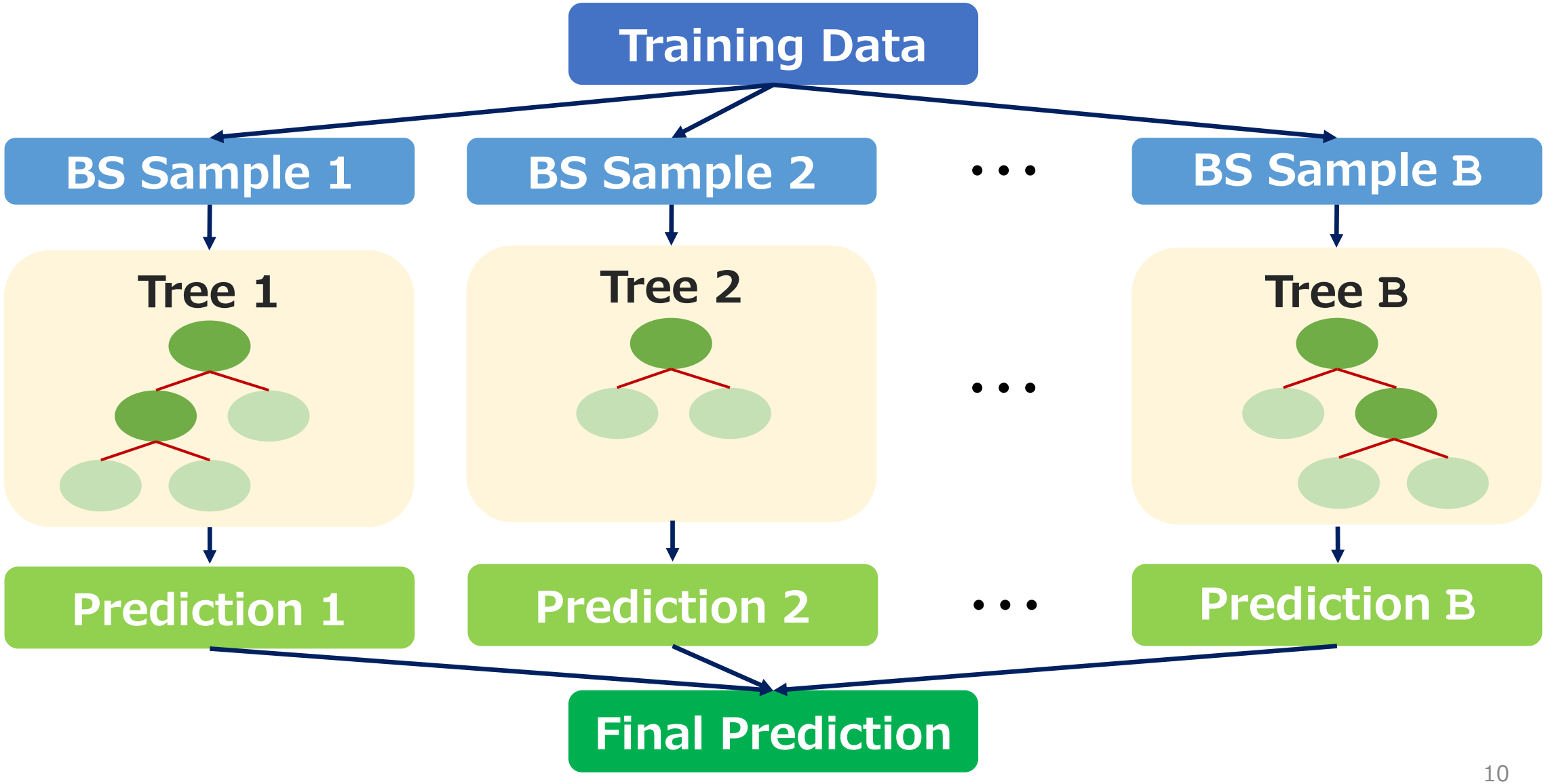
1. From the training data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, generate a data set (bootstrap sample) $\mathcal{D}_n^*$ by bootstrapping
2. With the dataset $\mathcal{D}_n^*$, generate a tree T^* where:
   - The pre-specified number `mtry` of features are randomly selected out ot p features for each branching
   - The branching is iterated until a sample size in each leaf reaches the pre-specified number `nodesize`
3. Step 1. & 2. are iterated `B` times, which generate `B` trees $T_1^*, \ldots, T_B^*$ (the set of them is called Random Forests)

Random Forests Algorithm

4. Get predictions for each tree for a new data
5. The final prediction is determined by:
   - Weighted average (regression);
   - Majority voting (classification) of predictions from all trees

Creating many trees that are less dependent on each other (diverse trees) and aggregating all tree can lead to a powerful predictor

Random Forests Image



In-Bag(IB) & Out-Of-Bag(OOB)

IB: a sample used for bootstrapping for a tree

OOB: a sample unused for bootstrapping for a tree

- circa 37% of train data would be OOB
- because the probability that a sample is not selected can be evaluated as follows:

$$\left(\frac{n-1}{n}\right)^{n} \approx \exp(-1) \approx 36.8\%$$

OOB prediction allows evaluation of generalization performance i.e. test errors can be evaluated without cross-validation

Feature Importance

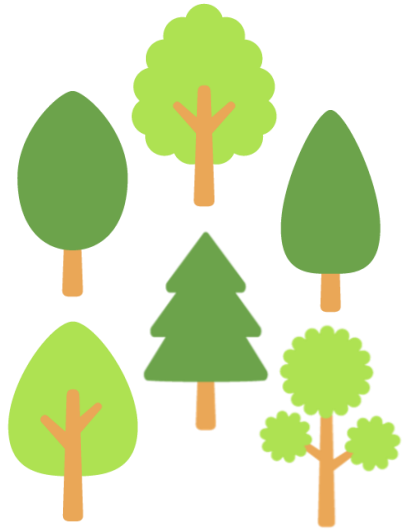
- Random forests combine plenty of decision trees, making it difficult to interpret the trees
- Instead, the importance of each explanatory variable (how much it contributes to the prediction) can be assessed
- Feature Importance is one of the typical method of IML: Interpretable Machine Learning

Feature Importance

Feature Importance could be assessed by the following indicators:

1. Amount of increase in error
   - The amount of increase in error when the explanatory variable to be evaluated is excluded

2. Increase in node impurity
   - Increase in impurity of all leaf nodes (residual sum of squares for regression, Gini coefficient for classification) when the explanatory variable to be evaluated is excluded.

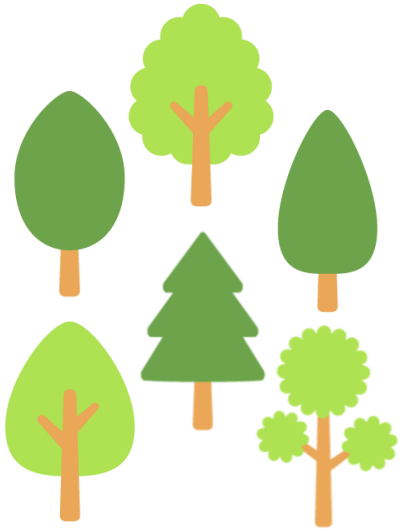


Properties Required for Actuarial Analysis

Requirements for Actuarial Analysis

- Justification & Explanation
  - ✓ Interpretability
  - ✓ Robustness
  - ✓ Statistical basis

- Extrapolation
  - ✓ Along time
  - ✓ Along risk factor values



Comparison with Other Methods

GLM

GLM has been familiar among actuaries for its high interpretability and robustness.

- Justification & Explanation
  - ✓ Robustness
  - ✓ High interpretability
  - Requires careful selection of polynomial & interaction terms to avoid bias
- Can extrapolate
- ➢ Relies too much on manually set assumptions & cannot make use of big data.

GAM

GAM has been used by actuaries for its simplicity and flexibility.

- Justification & Explanation
  - The curse of dimensionality with high dimension & high-order terms
  - ✓ Interpretability
  - ✓ Reduced bias with more flexibility than GLM with high-order terms
  - High bias remaining with only low-order terms
- Can extrapolate

➢ Dilemma of the curse of dimensionality or high bias

Neural Networks

NN has been known for its great performance given large amounts of data.

- Justification & Explanation
  - Results vary depending on tuning & gradient descent calculation
  - Difficult to interpret hidden layer variables
  - ✓ Good performance with large sample size
  - Poor performance with limited amount of data
- Can Extrapolate
- ➢ Low interpretability, poor performance & instability with limited sample size

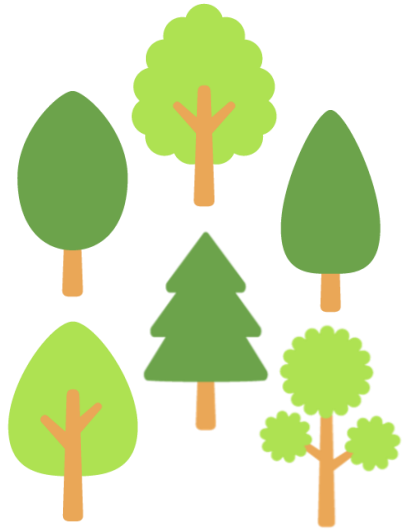
Gradient Boosting Machines

GBM has been widely used among data-scientists for its prediction accuracy.

- Justification & Explanation
  - Varying results depending on tuning & boosting
  - Very good accuracy with medium sample size
- Cannot extrapolate
- ➢ Unstable results depending on how tuning is done

Random Forests

- Random Forests are a classical algorithm proposed in 2001.
- Justification & Explanation
  - ✓ Easy tuning & relatively stable results
  - ✓ Good accuracy with medium sample size
  - ✓ Good statistical properties
    - ✓ Gives a prediction of Y's distribution, with consistency
    - ✓ Asymptotic normality (for GRF)
- Cannot extrapolate
- ➢ Stable results & statistical basis



Potential Use & Issues of RF in Actuarial Practice

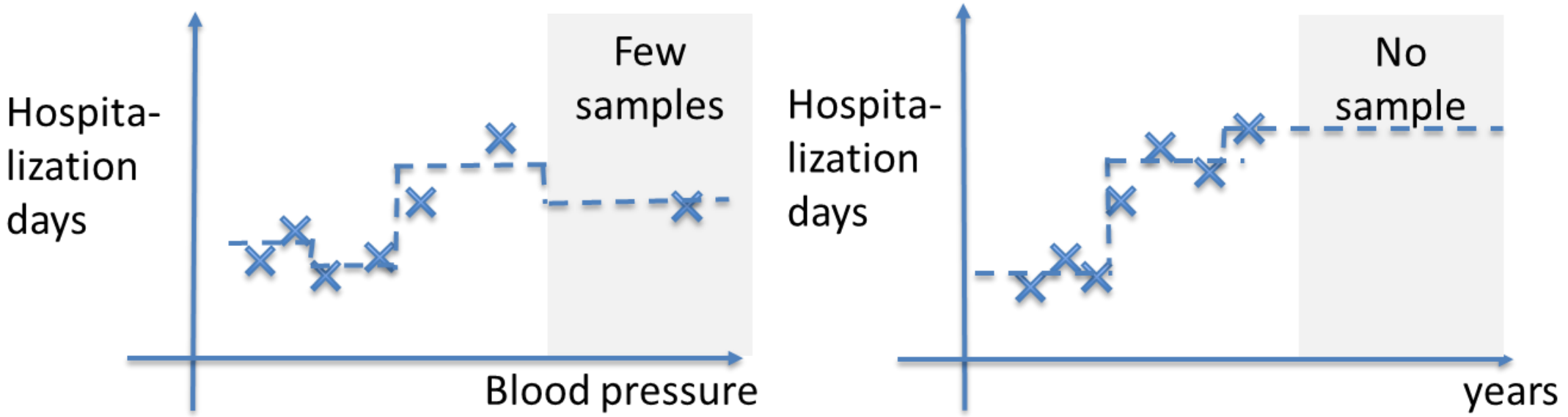
Use-cases of RF

- Known use cases
  - ✓EDA
  - ✓Baseline model

- Our proposed use cases
  - ✓Find a good threshold for risk segmentation
  - ✓Estimate the error distribution from a best-estimate prediction
    i.e. the distribution of $Y - E[Y|X = x]$
  - ✓Direct use for predictive analysis

Extrapolation Problem

- Tree-based models can only make local predictions, thus unable to be extrapolated to where few samples are available



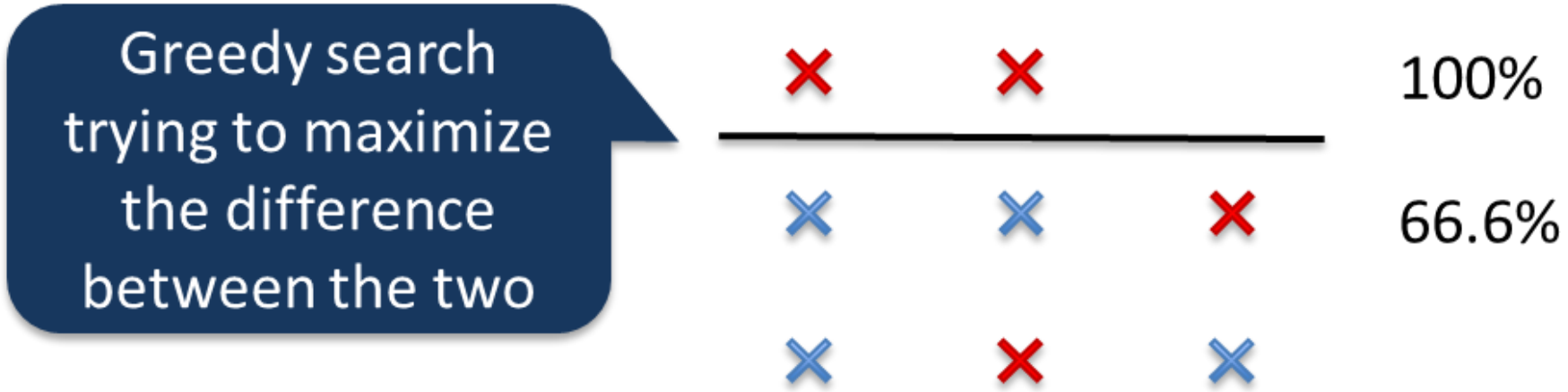
Improvement: Boosting with GLM

- To address the need for extrapolation, boosting RF with GLM or low-order GAM could be effective.
- We expect them to compensate for each other's weaknesses

	GLM/GAM	RF	GLM/GAM & RF
Extrapolation	○	×	○
Fit complex function & interaction	×	○	○
Error distribution	×	○	○

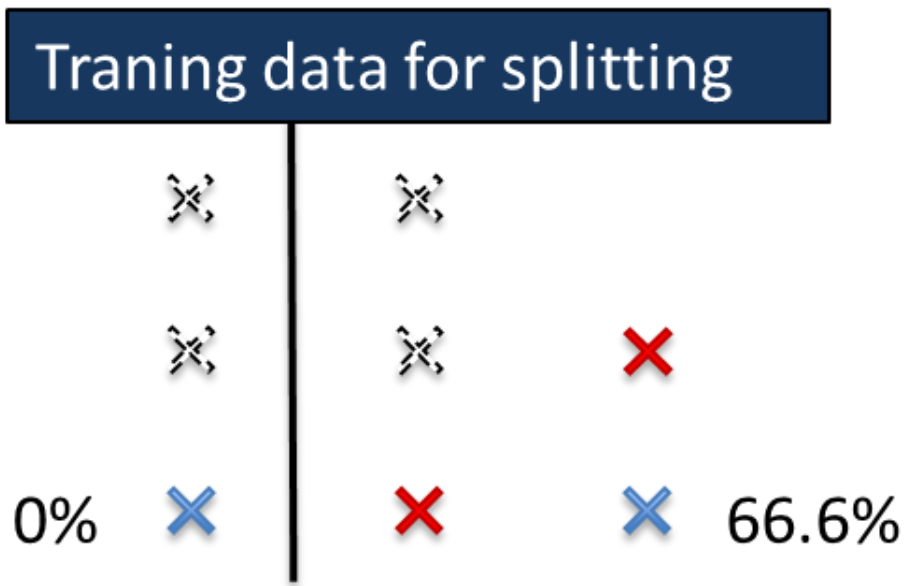
Noise Problem

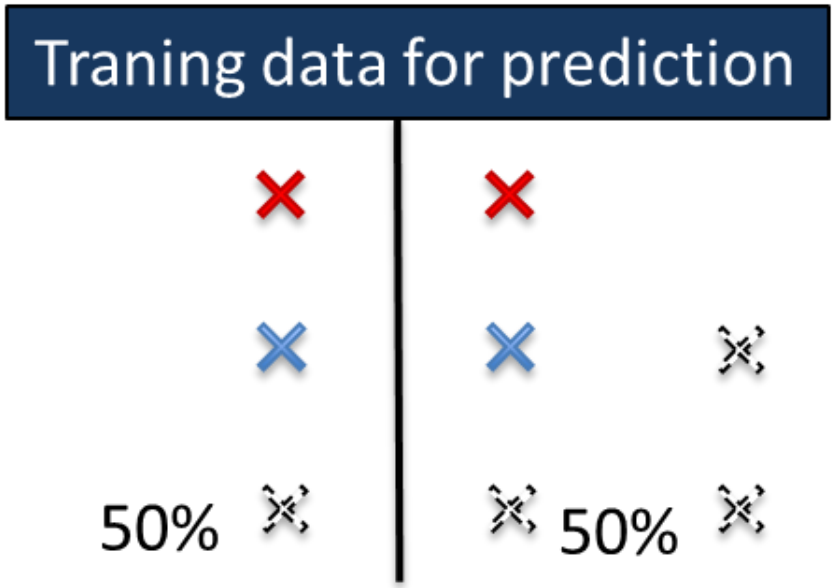
- RF can be overfit to local noise in training data, which causes unexplainable local noise in predictions.
  - One reason for this noise might be “not being honest”
  - Using the same samples to calculate predictions can result in positive bias in local difference of predictions.



Improvement: Honesty

- Honest trees/forests use separate training data for splitting and prediction.
- We expect them to have less noise, at the cost of accuracy due to smaller sample size.





Statistical Property: Asymptotic Normality

- Under some regularity conditions, asymptotic normality is shown for GRF, a generalized version RF with honesty.

$$\frac{\hat{\mu}_n^{RF}(x) - \mu(x)}{\sigma_n(x)} \to N(0,1) \qquad \text{for a sequence } \sigma_n \to 0$$

Reference

Honest trees

- Susan Athey, Guido Imbens.(2016). Recursive Partitioning for Heterogeneous Causal Effects

GRF

- Susan Athey, et al. (2019). Generalized Random Forests