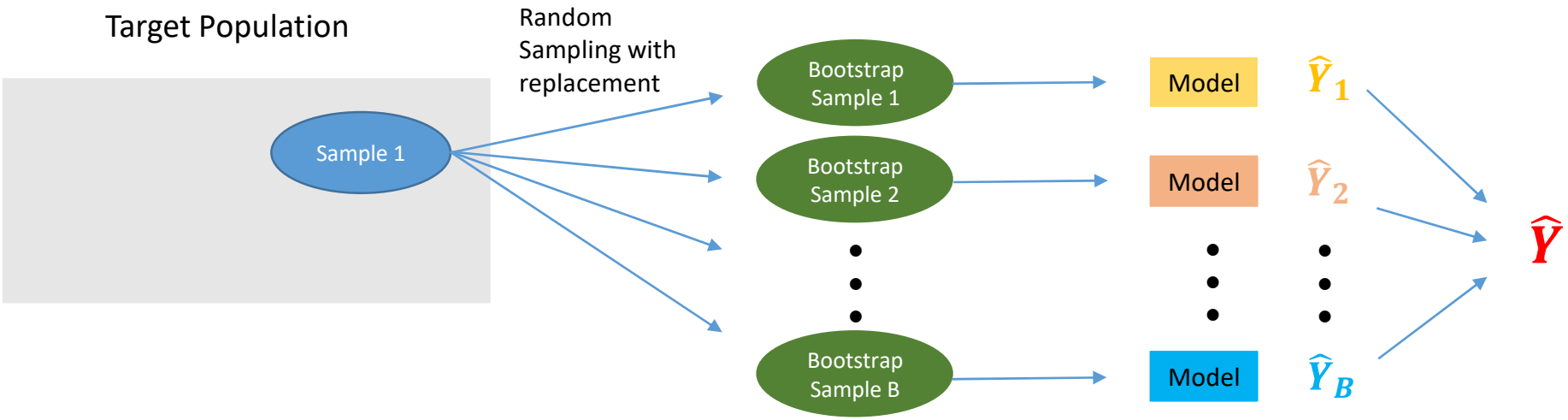
Random Forests

Part 2: Random Subset Feature (RSF) Selection

X variables

Recap: Bootstrap Aggregating (Bagging) Design



- Each bootstrap sample has the same size (n) as the original sample (n).
- Each of the bootstrap sample serves as a trainset for the same model type. e.g. CART.
  - Number of bootstrap samples = Number of models.
- B is chosen to be a large number. Breiman (1996a): 50 for Classification Tree, 25 for Regression Tree.
- Bagging Prediction:
  - If Y is continuous, take the mean of the B predictions Y_hat.
  - If Y is categorical, take the mode of the B predictions Y_hat. i.e. majority wins.
- Compute testset errors from OOB sample from each Bootstrap and take the mean.

Sources: Chew C.H. (2022) AI, Analytics and Data Science, Vol. 2, Chap. 8 & Breiman (1996a)

Results of Bagging Classification Trees (i.e. categorical Y) in Breiman (1996a)

Table 2. Misclassification Rates (%)

Data Set	$\bar{e}_S$	$\bar{e}_B$	Decrease
waveform	29.1	19.3	34%
heart	4.9	2.8	43%
breast cancer	5.9	3.7	37%
ionosphere	11.2	7.9	29%
diabetes	25.3	23.9	6%
glass	30.4	23.6	22%
soybean	8.6	6.8	21%

Table 3. Standard Errors of Misclassification

Data Set	$SE(\bar{e}_S)$	$SE(\bar{e}_B)$
waveform	.2	.1
heart	.2	.1
breast cancer	.3	.2
ionosphere	.5	.4
diabetes	.4	.4
glass	1.1	.9
soybean	.4	.3

- Testset errors substantially decreased with Bagging.
- Standard errors decreased too.

Source: Breiman (1996a)

Results of Bagging Regression Trees (i.e. continuous Y) in Breiman (1996a)

Table 8. Mean Squared Test Set Error

Data Set	$\bar{e}_S$	$\bar{e}_B$	Decrease
Boston Housing	20.0	11.6	42%
Ozone	23.9	18.8	21%
Friedman #1	11.4	6.1	46%
Friedman #2	31,100	22,100	29%
Friedman #3	.0403	.0242	40%

Table 9. Standard Errors

Data Set	$SE(\bar{e}_S)$	$SE(\bar{e}_B)$
Boston Housing	1.0	.6
Ozone	.8	.6
Friedman #1	.10	.06
Friedman #2	300	100
Friedman #3	.0005	.0003

- Testset errors substantially decreased with Bagging.
- Standard errors decreased too.

Source: Breiman (1996a)

But Bagging results are not as significant for Bagging KNN models

Table 11. Misclassification Rates for Nearest Neighbor

Data Set	$\bar{e}_S$	$\bar{e}_B$
waveform	26.1	26.1
heart	5.1	5.1
breast cancer	4.4	4.4
ionosphere	36.5	36.5
diabetes	29.3	29.3
glass	30.1	30.1

- Testset errors are almost the same with or without Bagging.
- KNN models are relatively stable models.

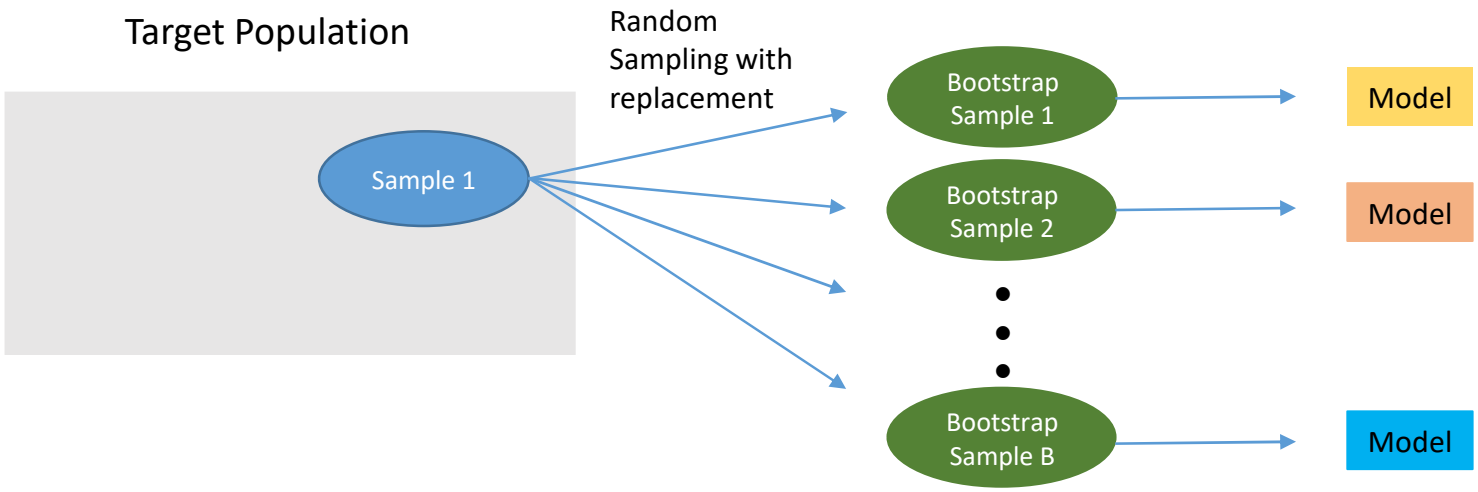
Source: Breiman (1996a)

Stability of the model

- Refers to the stability of the procedure in specifying the model.
  - Eg: Getting the model coefficients in linear reg via min. SSE.
- Improvement will occur [in Bagging] for unstable procedures where a small change in training set can result in large changes in the model.
- Examples of unstable procedures:
  - Neural Network
  - CART
  - MARS
  - Subset selection in Linear Regression
- Examples of stable procedures:
  - KNN
  - Ridge Regression

Key Learning about Stability of Models

- Bagging is effective for unstable procedures – stabilizing effect.
- Bagging will be ineffective for stable procedures.
- But why?



Source: Chew C.H. (2022) A.I., Analytics & Data Science, Vol. 2.

Key Learning about Stability of Models

- Bagging is effective for unstable procedures – stabilizing effect.
- Bagging will be ineffective for stable procedures.
- But why?

Model 1 ≈ Model 2 ≈ ... ≈ Model B

If all the models are almost the same, then there is no benefit of Bagging.

Stability of a procedure is relative

- Linear Regression is a unstable procedure.
- In general, linear Regression with all Xs [no selection] is more stable than Linear reg with selected Xs.
- i.e. If you have to select the Xs into Linear Regression, procedure becomes more unstable.

Effect of Dominant X on Stability

- Example:
  - Predict Y = Course Grade,
  - X_1 = Hours studying,
  - X_2 = Statistics Exam Marks,
  - X_3 = Course Exam Marks (70% of Grade).
- X_3 is dominant in predicting Y.
- Any model that predicts Y will be using dominant X_3.
- Various Bootstrap models become more similar and hence, could not significantly improve over single model.

Two Reasons Why Bagging is not Useful

1. Stable Procedures.
2. Dominant X(s).

- Reason: Different bootstrap samples producing the same or almost the same model.
- How to improve Bagging results?
  - How to infuse controlled instability, even in the presence of dominant X?
    - Risk in using a more unstable procedure: errors might increase.
    - Controlled Instability: Maintain accuracy while increasing diversity of models.
    - Ans:

    1. Random Subset Feature Selection (Different subset of Xs, randomly at each split), and at the same time,
    2. Grow the CART tree to the max from each bootstrap sample to maintain high accuracy (no pruning). The bigger the tree, the higher the accuracy on trainset.

Role of Random Subset Feature(RSF) Selection

- At each split of a tree, a randomly chosen subset of the X variables are considered and searched to determine the best split.
- i.e. the non-chosen X variables have no chance to be selected as the best split.
- At another split, another randomly chosen subset of the X variables are considered and searched to determine the best split.
- Effectively:
  - Infuses Instability, as another tree would have a different random subset of Xs available to consider.
  - Prevents any dominant X(s) from dominating all the way to produce the same model.

Random Forest Results in Breiman (2001)

Selection: Best of RSF = 1 and RSF = int($\log_2$(M)+1)

Table 2. Test set errors (%).

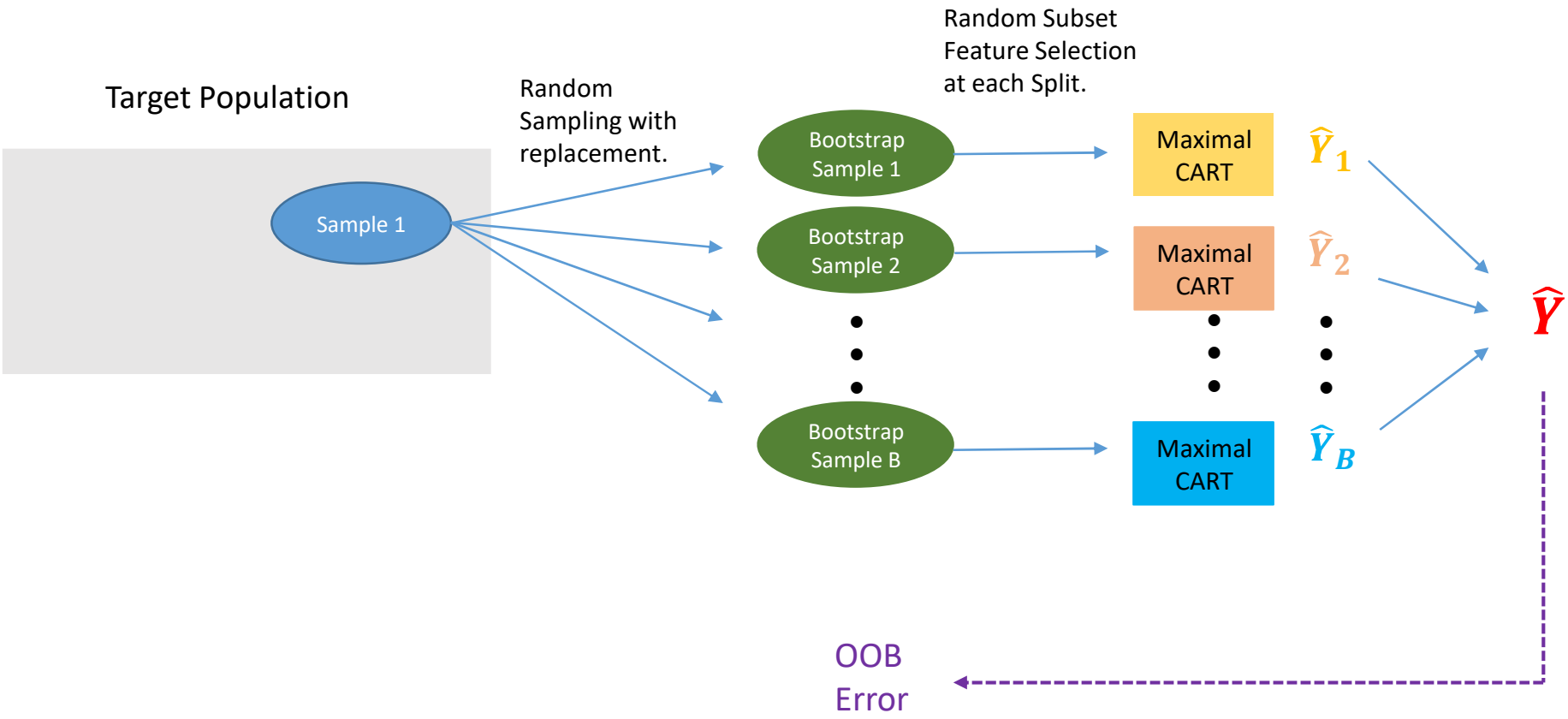
Data set	Adaboost	Selection	Forest-RI single input	One tree
Glass	22.0	20.6	21.2	36.9
Breast cancer	3.2	2.9	2.7	6.3
Diabetes	26.6	24.2	24.3	33.1
Sonar	15.6	15.9	18.0	31.7
Vowel	4.1	3.4	3.3	30.4
Ionosphere	6.4	7.1	7.5	12.7
Vehicle	23.2	25.8	26.4	33.1
German credit	23.5	24.4	26.2	33.3
Image	1.6	2.1	2.7	6.4
Ecoli	14.8	12.8	13.0	24.5
Votes	4.8	4.1	4.6	7.4
Liver	30.7	25.1	24.7	40.6
Letters	3.4	3.5	4.7	19.8
Sat-images	8.8	8.6	10.5	17.2
Zip-code	6.2	6.3	7.8	20.6
Waveform	17.8	17.2	17.3	34.0
Twonorm	4.9	3.9	3.9	24.7
Threenorm	18.8	17.5	17.5	38.4
Ringnorm	6.9	4.9	4.9	25.7

Source: Breiman (2001)

Random Forest = Bagging + Random Subset Feature

Source: Chew C.H. (2022) A.I., Analytics & Data Science, Vol. 2.

Random Forest Process



Next – Random Forest with R

- What are the default settings for B and RSF size in R?
- Where do we see OOB error?
- How do we check if the error had converged?

  "…in practice we use a value of B sufficiently large for the error rate to have settled down." -- *ISLR*

- What are the useful charts and diagnostics in R?
- What to watch out for if we use Python implementation instead of R?