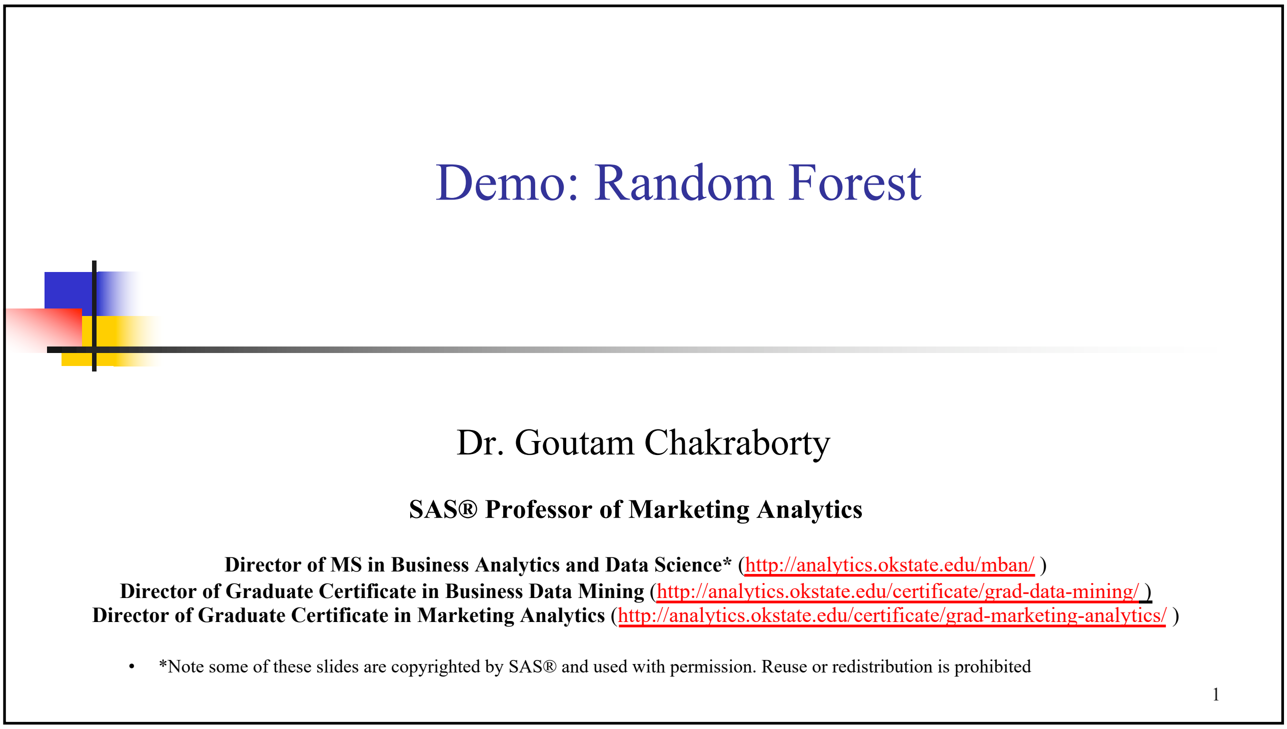


Demo: Random Forest

Dr. Goutam Chakraborty

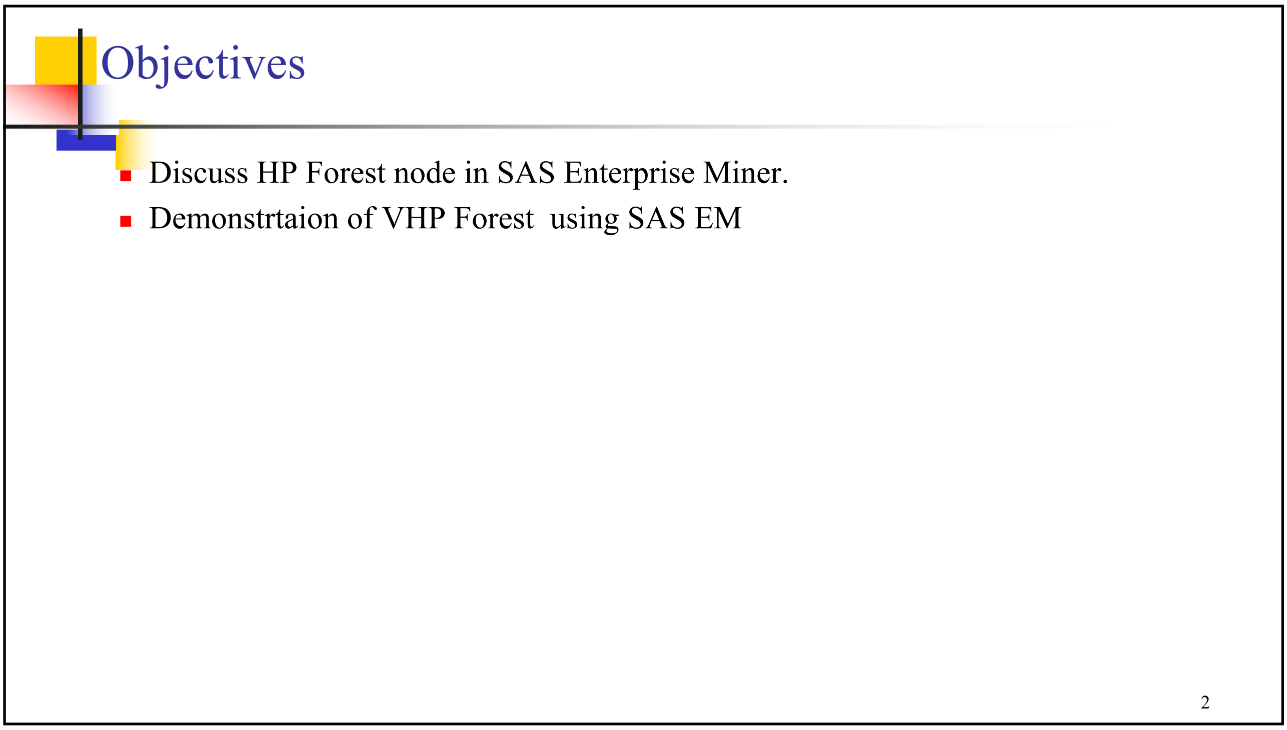
SAS® Professor of Marketing Analytics

Director of MS in Business Analytics and Data Science* (http://analytics.okstate.edu/mban/)
Director of Graduate Certificate in Business Data Mining (http://analytics.okstate.edu/certificate/grad-data-mining/)
Director of Graduate Certificate in Marketing Analytics (http://analytics.okstate.edu/certificate/grad-marketing-analytics/)

- *Note some of these slides are copyrighted by SAS® and used with permission. Reuse or redistribution is prohibited

1

1



Objectives

- Discuss HP Forest node in SAS Enterprise Miner.
- Demonstrtaion of VHP Forest using SAS EM

2

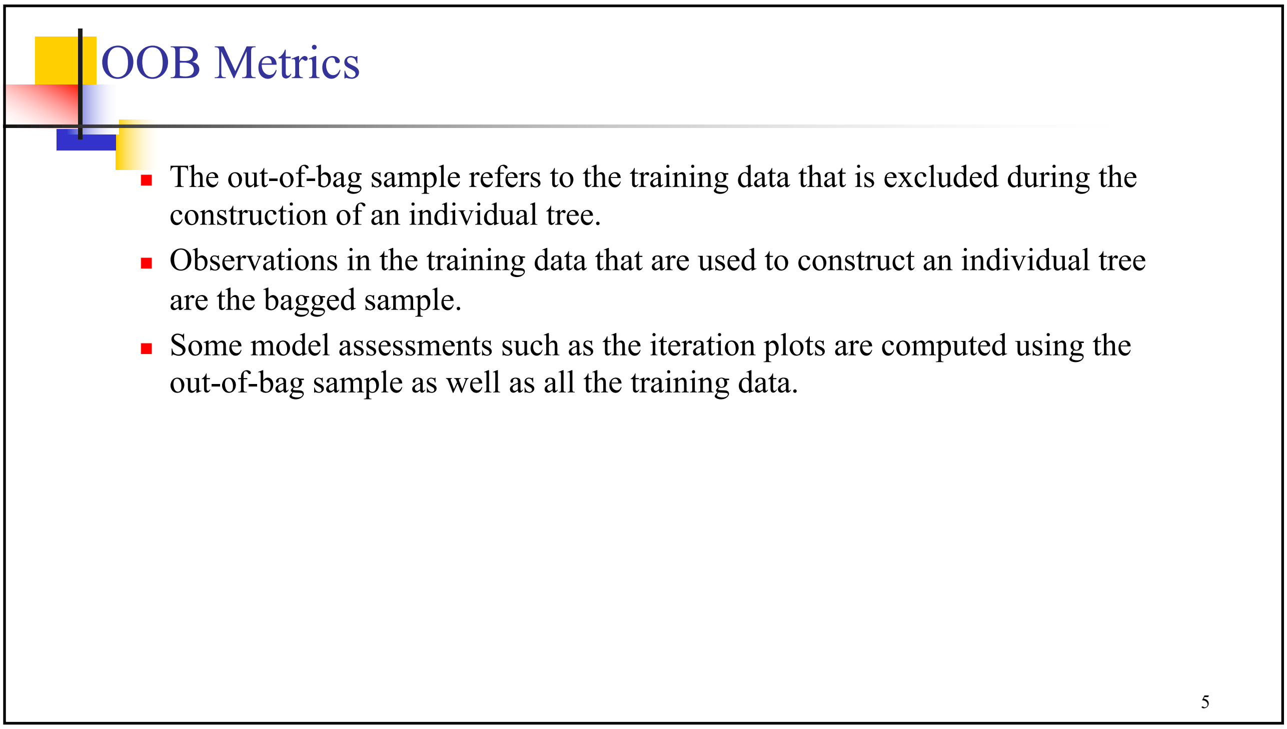
2

HP Forest Algorithm

- *Bagging* is the term for averaging many trees grown on bootstrap samples of the rows of training data. All columns are considered for splitting at every step.
- The HP Forest algorithm in SAS EM does sampling of the rows ***and*** sampling of the columns at each step.
- The forest algorithm perturbs the training data more than the bagging algorithm, producing more variation among the trees in the ensemble.
- Ensembles of a more diverse set of trees often leads to improved predictive accuracy.

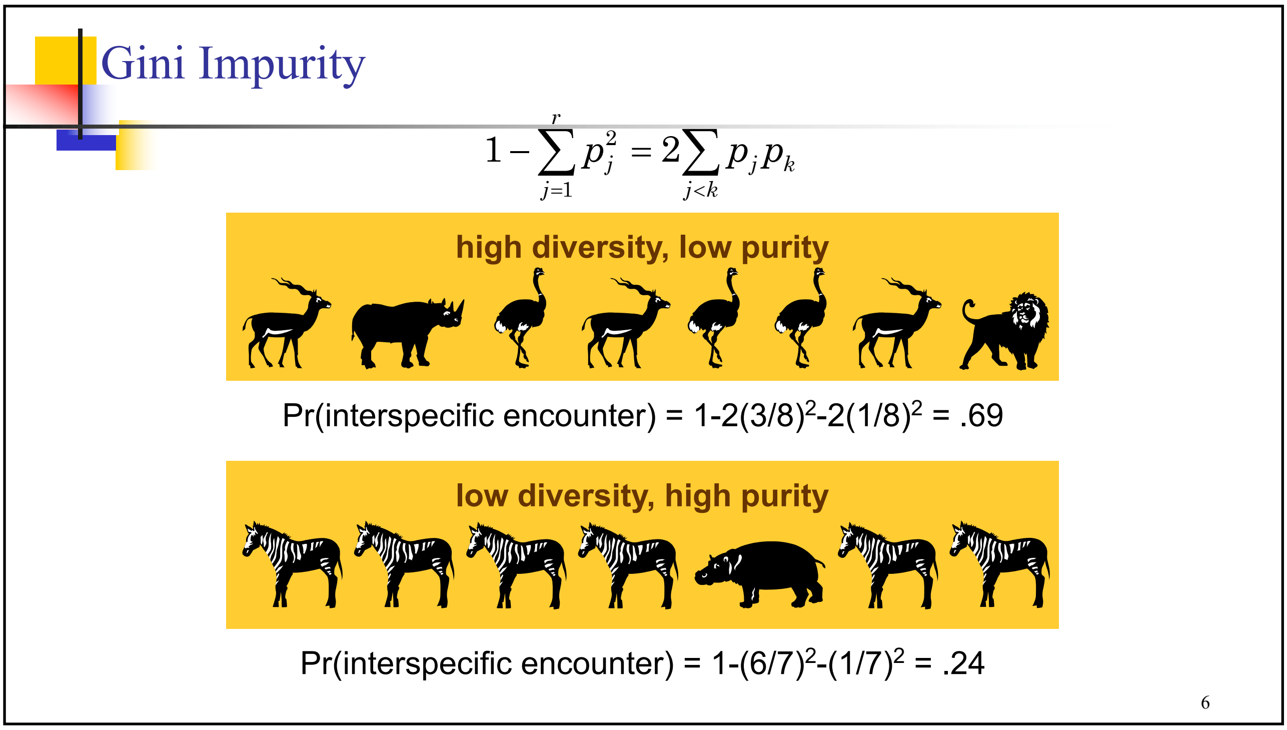
HP Forest Node

- These are the three main options:
  - Number of trees
    - Specifies the number of trees that make up the forest. (Default = 100)
  - Number of inputs for a node
    - Specifies the number of input variables to consider splitting for each node. (Default = $\sqrt{\#\ of\ inputs}$)
  - Sampling strategy
    - Specifies the number of observations used for each tree and how this sample is obtained.
      (Default = “proportion” and 0.6)



OOB Metrics

- The out-of-bag sample refers to the training data that is excluded during the construction of an individual tree.
- Observations in the training data that are used to construct an individual tree are the bagged sample.
- Some model assessments such as the iteration plots are computed using the out-of-bag sample as well as all the training data.

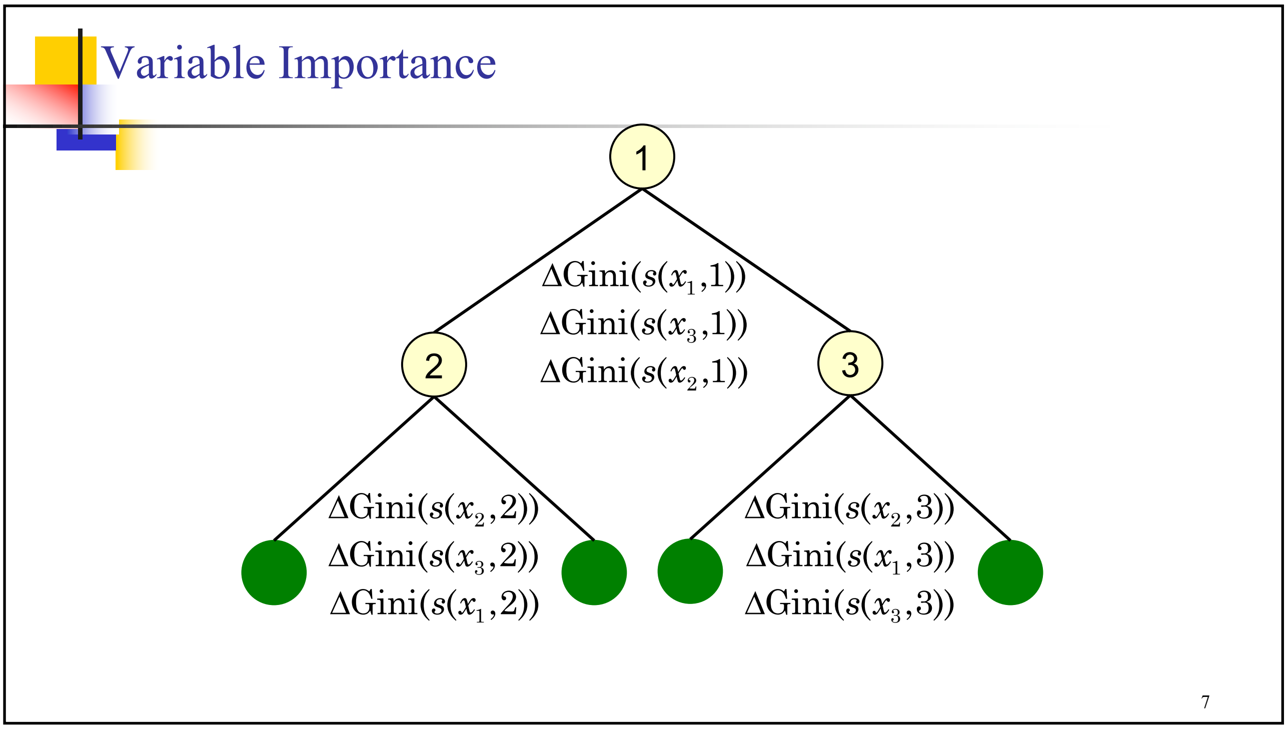


Gini Impurity

$$1-\sum_{j=1}^{r} p_j^2 = 2\sum_{j<k} p_j p_k$$

Pr(interspecific encounter) = 1-2(3/8)^2-2(1/8)^2 = .69

Pr(interspecific encounter) = 1-(6/7)^2-(1/7)^2 = .24



Variable Importance

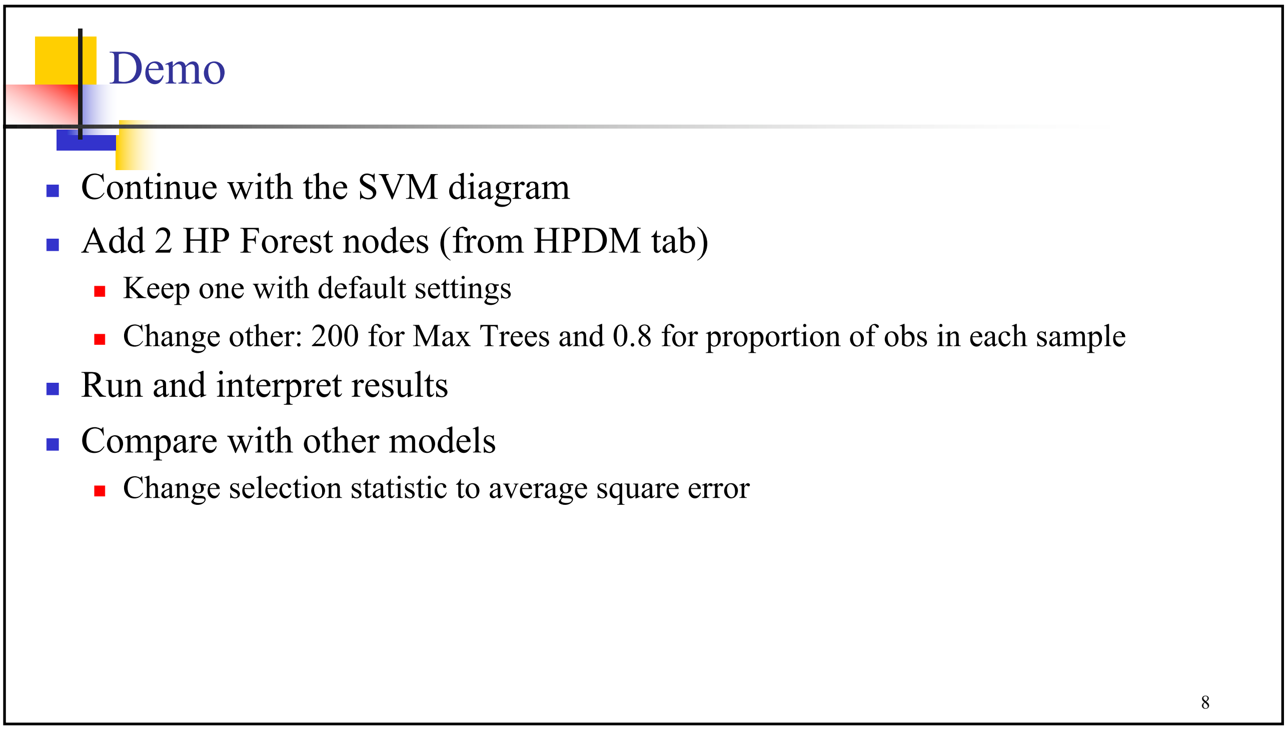
$\Delta\text{Gini}(s(x_1,1))$
$\Delta\text{Gini}(s(x_3,1))$
$\Delta\text{Gini}(s(x_2,1))$

$\Delta\text{Gini}(s(x_2,2))$
$\Delta\text{Gini}(s(x_3,2))$
$\Delta\text{Gini}(s(x_1,2))$

$\Delta\text{Gini}(s(x_2,3))$
$\Delta\text{Gini}(s(x_1,3))$
$\Delta\text{Gini}(s(x_3,3))$

7

7



Demo

- Continue with the SVM diagram
- Add 2 HP Forest nodes (from HPDM tab)
  - Keep one with default settings
  - Change other: 200 for Max Trees and 0.8 for proportion of obs in each sample
- Run and interpret results
- Compare with other models
  - Change selection statistic to average square error

8

8