

An Introduction to Tree-Based Methods

Xiaogang Su, Ph.D.
Department of Statistics
University of Central Florida
Orlando, FL 32816

Outline

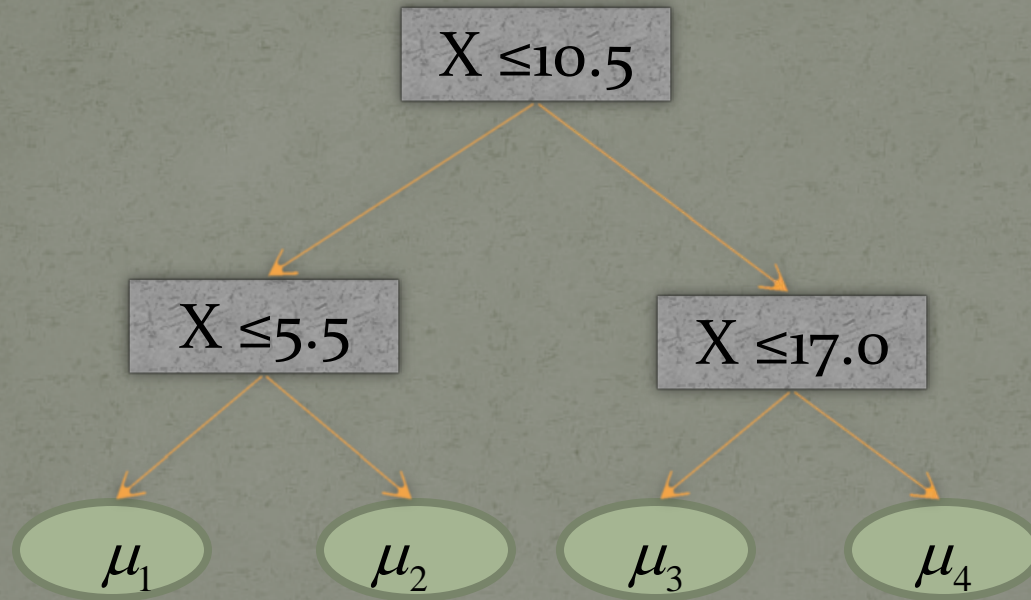
1. What is a tree-structured model?
2. History of tree modeling
3. Why tree-based methods?
4. The current standard of constructing trees – CART methodology
 - ◆ Growing a large initial tree
 - ◆ Pruning
 - ◆ Tree size selection via validation
5. Implementation
6. An Example

1, What is a Tree Model?

A tree-based method (or *recursive partitioning*) recursively partitions the predictor space to model the relationship between two sets of variables (responses vs. predictors).

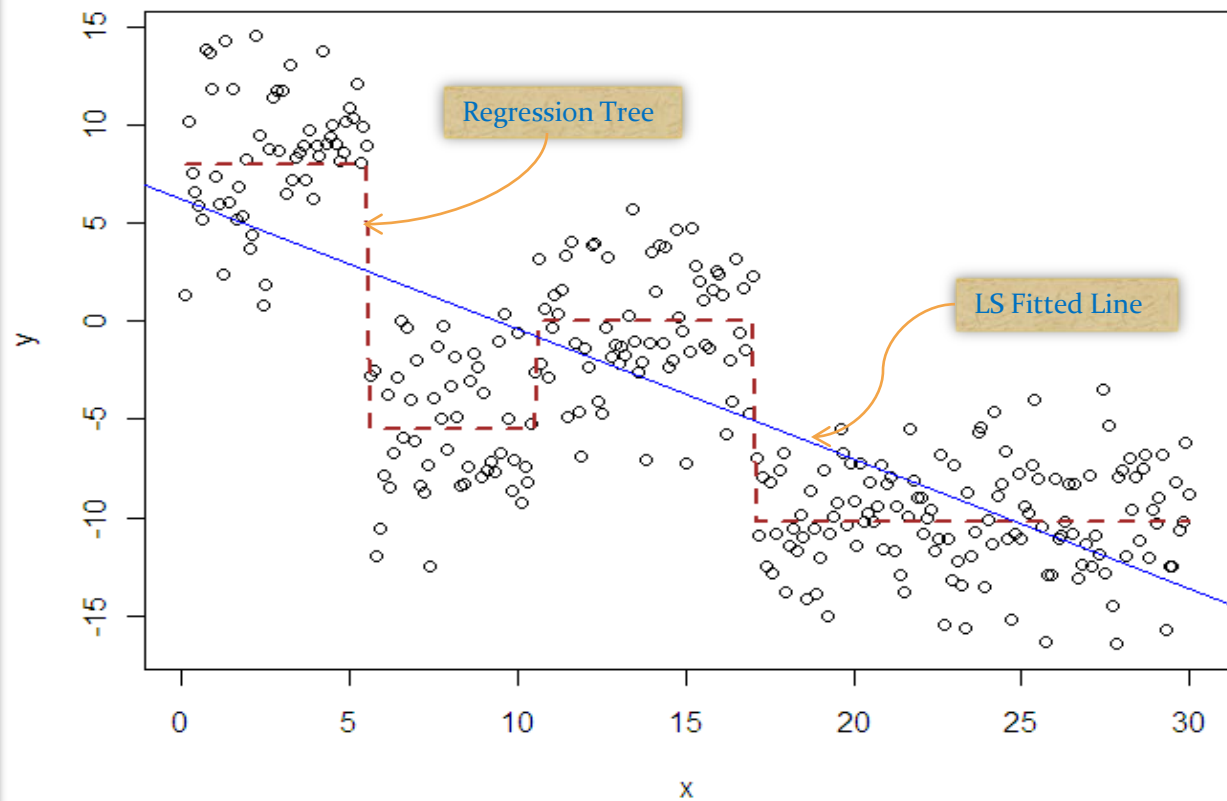
- Tree modeling facilitates piecewise constant fitting.
- Two Types in Data Mining:
 - Regression Trees (continuous response)
 - Classification /Decision Trees (categorical response).

A Simple Regression Example

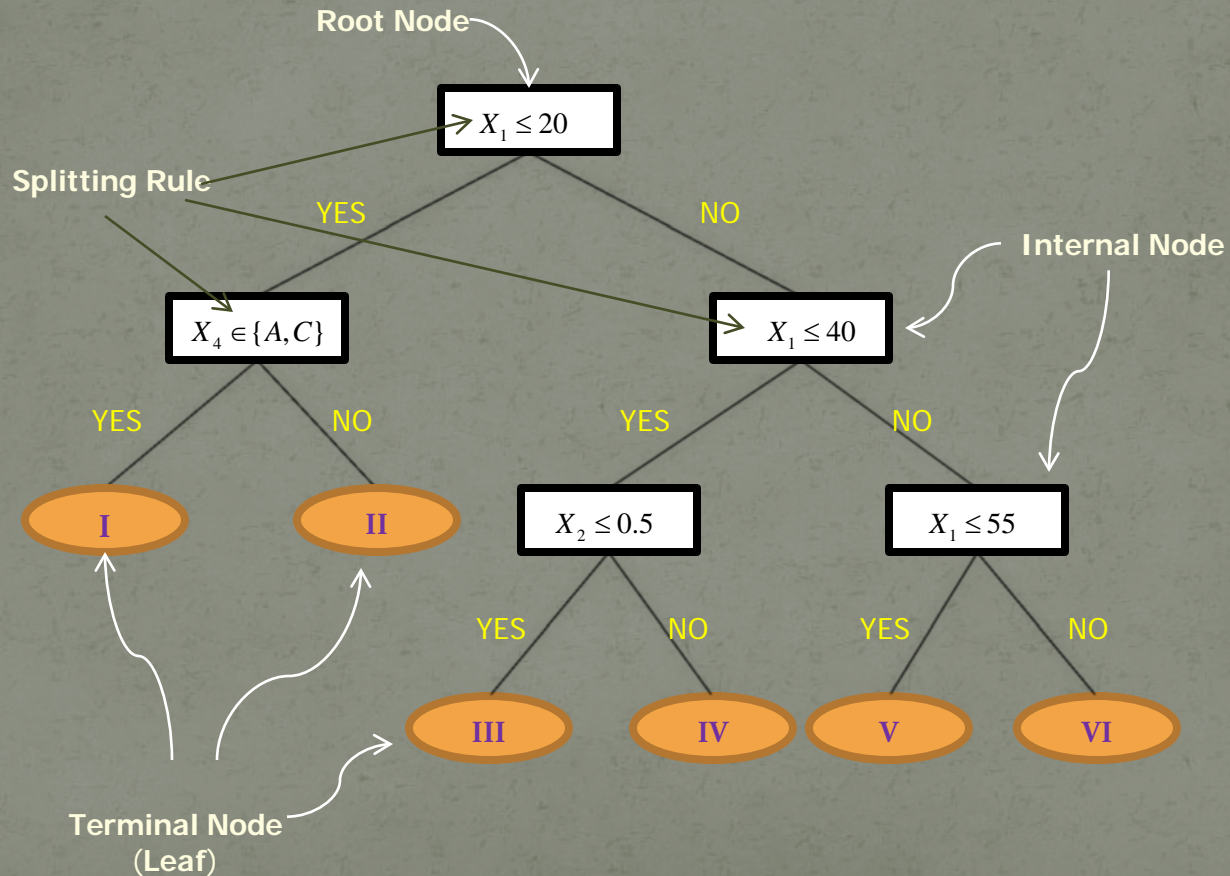


$$y = \mu_1 \cdot 1_{\{x \leq 5.5\}} + \mu_2 \cdot 1_{\{x > 5.5\}} \cdot 1_{\{x > 10.5\}} + \mu_3 \cdot 1_{\{x > 10.5\}} \cdot 1_{\{x \leq 17\}} + \mu_4 \cdot 1_{\{x > 17\}}$$

An Illustration



Some Tree Terminology



Outline

1. What is a tree-structured model?
2. **History of tree modeling**
3. Why tree-based methods?
4. The current standard of constructing trees - CART
 - ◆ Growing a large initial tree
 - ◆ Pruning
 - ◆ Tree size selection via validation
5. Implementation
6. An Example

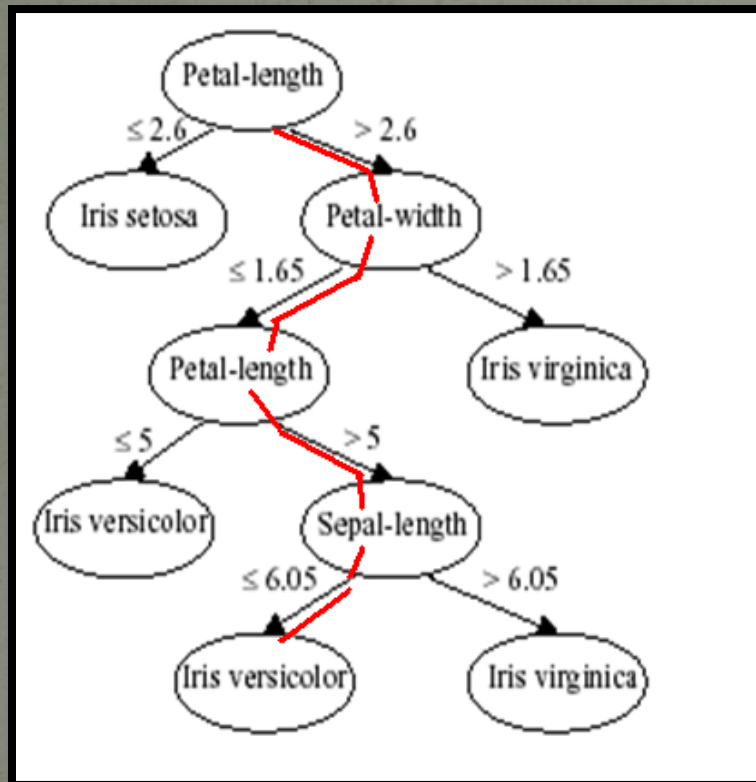
A Brief History

- Morgan and Sonquist (1963, *JASA*) –Automatic Interaction Detection (AID)
 - Hartigan (1975) and Kass (1980) - CHAID
- Breiman, Friedman, Olshen and Stone (1984) – Classification And Regression Trees (CART)
 - Another Similar Implementation - Quinlan (1979) - ID3 & C4.5
- Extensions
 - Multivariate Adaptive Regression Splines (MARS) – Friedman (1991);
 - Bagging (Breiman, 1996);
 - Boosting (Freund and Schapire (1996) and Friedman (2001)
 - Random Forests - Breiman (2003)

Outline

1. What is a tree-structured model?
2. History of tree modeling
3. Why tree-based methods?
4. The current standard of constructing trees - CART
 - ◆ Growing a large initial tree
 - ◆ Pruning
 - ◆ Tree size selection via validation
5. Implementation
6. An Example

Why Decision Trees?



IF

Petal-length > 2.6 AND
Petal-width ≤ 1.65 AND
Petal-length > 5 AND
Sepal-length > 6.05

THEN

the flower is Iris virginica

This is not the only rule for this species. What is the other?

Figure adapted from [SGI2001]

Interpretability

Example: When a bank rejects a credit card application, it is better to explain to the customer that it was due to the fact that

- He/she is not a permanent resident of Australia **AND**
- He/she has been residing in Australia for < 6 months **AND**
- He/she does not have a permanent job.

◆ This is better than saying:

“We are very sorry, but our neural network thinks that you are not a credit-worthy customer.” (In which case the customer might become angry and move to another bank)

◆ Interpretability is one of the main advantages of decision trees.

Pros

- It is nonparametric requiring few statistical assumptions and hence robust.
- It ameliorates the “curse of dimensionality” (a term due to Richard Bellman) and can be applied to various data structures involving both ordered and categorical variables in a simple and natural way.
 - In particular, the recursive partitioning method is exceptionally efficient in handling categorical predictors.

Pros (Continued)

- It does variable selection, complexity reduction, and (implicit) interaction handling in an automatic manner.
- Invariant under all monotone transformations of individual ordered predictors
- The output gives easily understood and interpreted information.
- Special features are available in handling missing values and obtaining ranking of variables in terms of their importance.

Cons

- Sometimes, not doing so well for estimation or prediction tasks.
- Instability of tree models

These two weaknesses can be utilized and improved by MARS, bagging, boosting, and random forests (Perturb & Ensemble procedures).

Outline

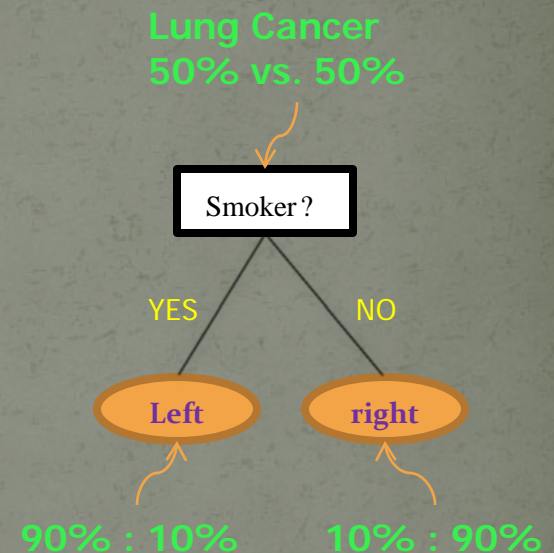
1. What is a tree-structured model?
2. History of tree modeling
3. Why tree-based methods?
4. **CART methodology of constructing trees**
 - ◆ Growing a large initial tree
 - ◆ Pruning
 - ◆ Tree size selection via validation
5. Implementation
6. An Example

Building a Tree Model

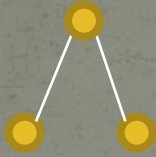
1. How to split data or grow trees?
2. How to declare a terminal node and determine the optimal tree size?
 1. How to summarize each terminal node?

Splitting

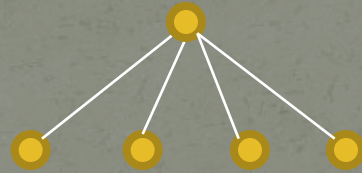
- A split is preferable if observations within the resultant child nodes become more homogeneous or less impure in terms of the response values or classes.
 - Impurity Measures: misclassification rate, entropy, Gini index
- Equivalently, after the split, observations between the two resultant child nodes become more heterogeneous.
 - Any appropriate two-sample test statistic can be used as such a measure of Goodness of Split.
- Maximize the measure over all permissible splits to identify the best one.



Why binary splits?



A binary split



A multi-way split

- A multi-way split can always be obtained by applying several binary split.
- It is hard to achieve optimality with multi-way splits due to increases number of allowable splits and difficulty in comparing across multi-way splits.

Several Other Issues

- Split on a categorical variable
 - 'Ordinal'ize the variable first when it has too many levels
- Select the splitting variable first and then apply greedy search to identify the best cutoff point
- Split with random sampling
- Stop splitting till a large initial tree is obtained.

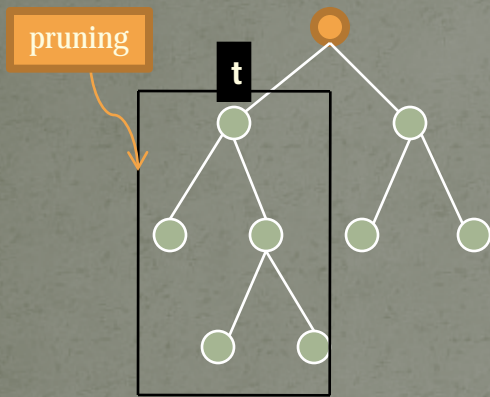
Pruning

- Stopping Rules is problematic.
 - Underfitting or Overfitting
- Pruning - CART
 - First growing a large initial tree and one of its subtrees is to be selected as the best tree structure.
 - To narrow down the choice of subtrees, a pruning algorithm is applied to iteratively truncate the initial large tree into a sequence of nested subtrees.
 - Test sample or v -fold cross-validation is then used to select the best subtree or determine the best tree size.

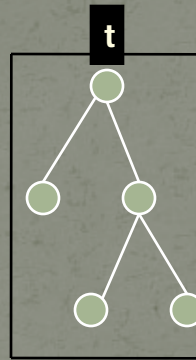
Pruning

- Any subtree of the initial tree could be the final tree model.
 - A tree T_1 is called a subtree of tree T , if every node of T_1 is included in T .
- However, the number of subtrees increases rapidly with the size of the initial tree. To see this, number of subtrees are 1, 2, 5, 26, 677, 458330 ... for full trees of depth 1, 2, ...
- Breiman et al. (1984) proposed an efficient pruning algorithm to help lower down the number of candidate subtrees.

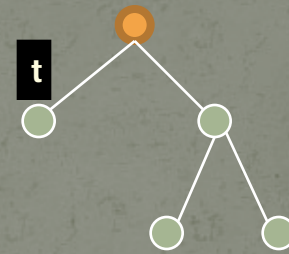
A Diagram for Pruning



(a) tree



(b) branch



(c) pruned subtree

Tree Size Determination

- The best tree is to be selected, via some validation method, from the *subtree sequence* obtained from the pruning algorithm.
- Validation Methods
 - Test Sample
 - Cross-Validation or Bootstrap resampling techniques
- Selection Criteria
 - GCV, AIC, BIC, etc.

Outline

1. What is a tree-structured model?
2. History of tree modeling
3. Why tree-based methods?
4. CART methodology of constructing trees
 - ◆ Growing a large initial tree
 - ◆ Pruning
 - ◆ Tree size selection via validation
5. **Implementation**
6. An Example

Implementation in SAS and R

- Two R Packages Available
 - **tree**
 - Functions: `tree()`, `prune.tree()`, `cv.tree()`, `predict.tree()`
 - **Rpart**
 - Functions: `rpart()`, `prune.rpart()`, `post.rpart()`
 - **Party – similar to CHAID**
 - Functions: `ctree()`
- SAS Implementation
 - SAS Enterprise Miner
 - PROC SPLIT

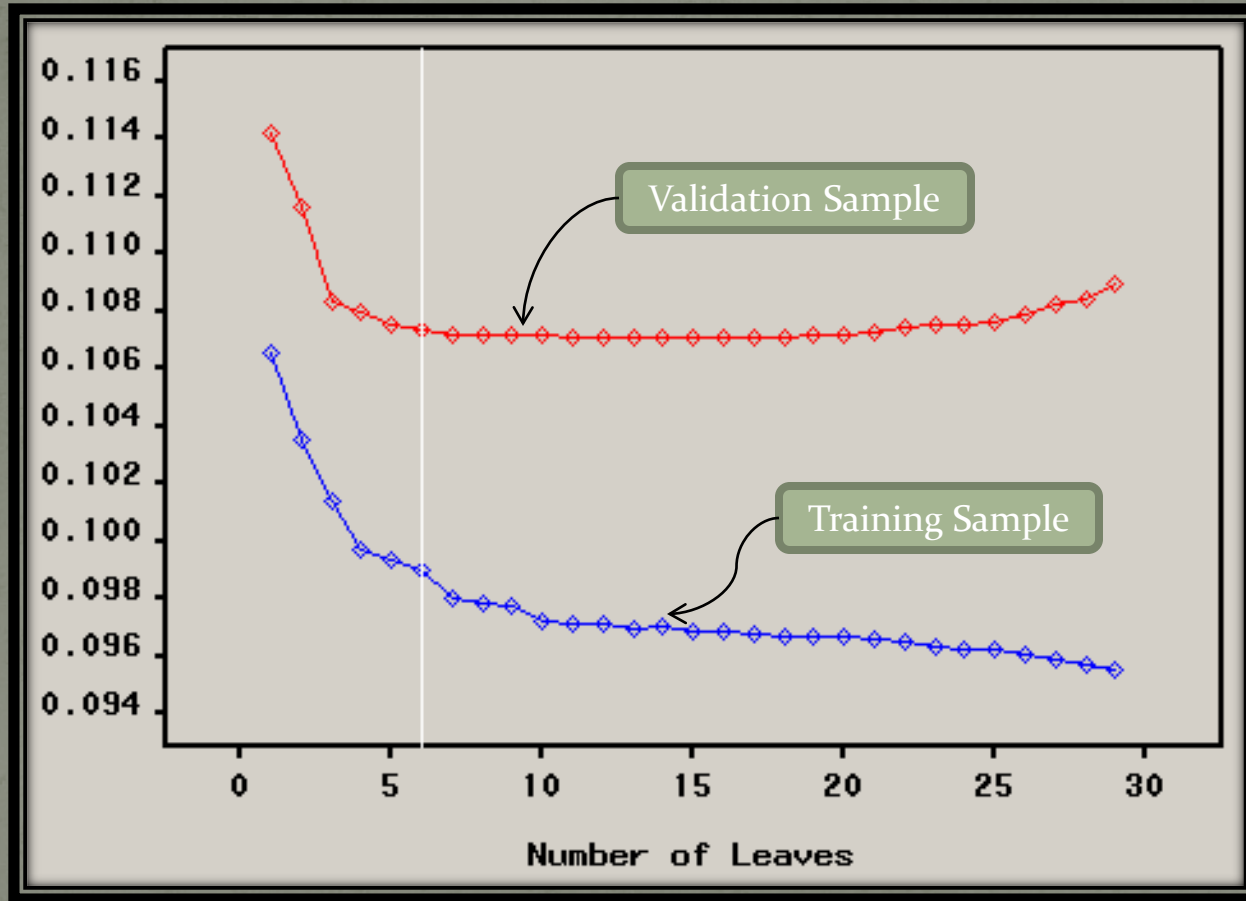
Outline

1. What is a tree-structured model?
2. History of tree modeling
3. Why tree-based methods?
4. The current standard of constructing trees – CART methodology
 - ◆ Growing a large initial tree
 - ◆ Pruning
 - ◆ Tree size selection via validation
5. Implementation
6. An Example

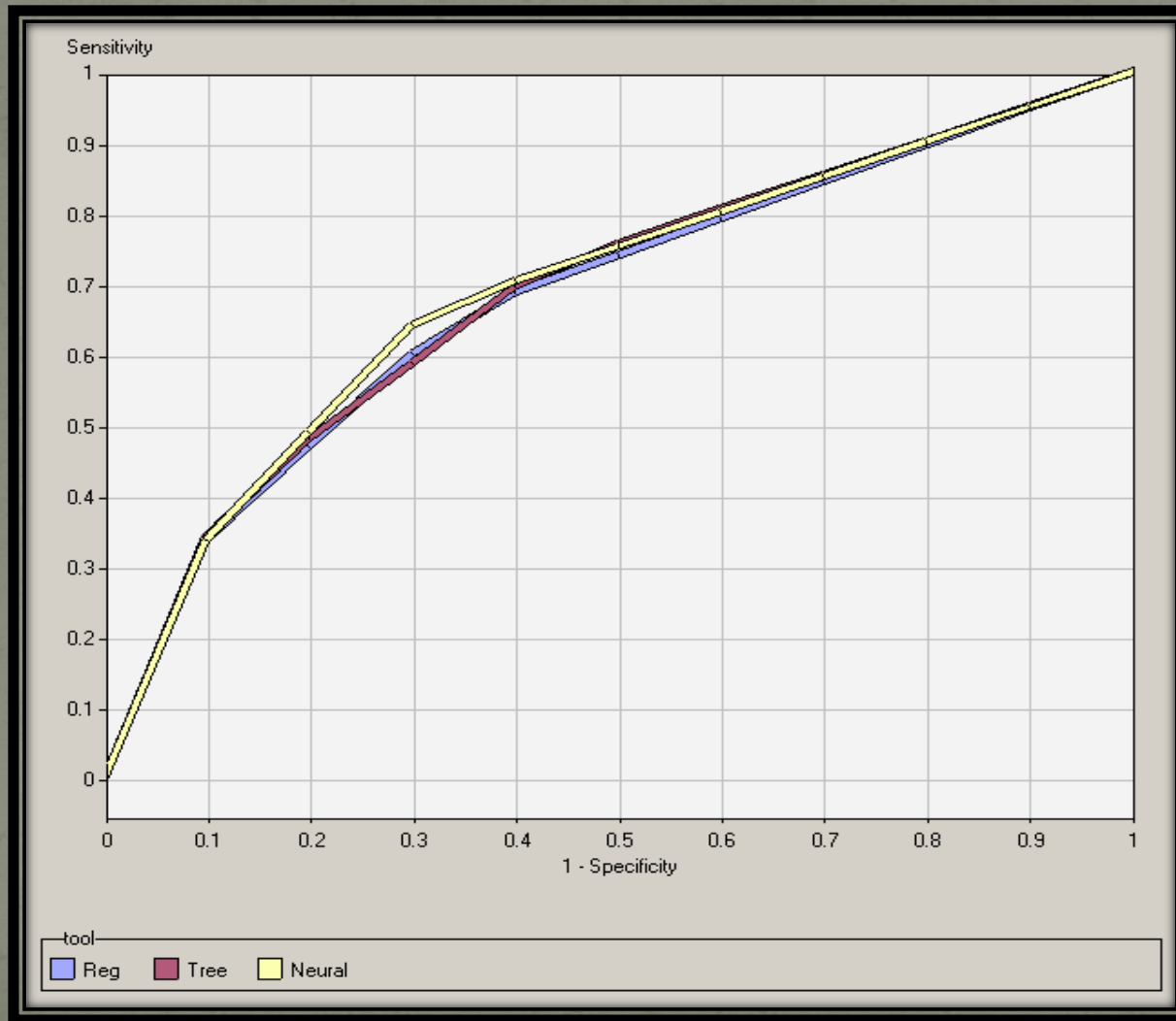
The Hospitalization Data

- Obtained from a Healthcare Consulting Company
- $N = 24,935$ and $p=83$ predictors.
- To predict whether an insured has hospitalization in the following year based on his/her demographics and previous-year info.

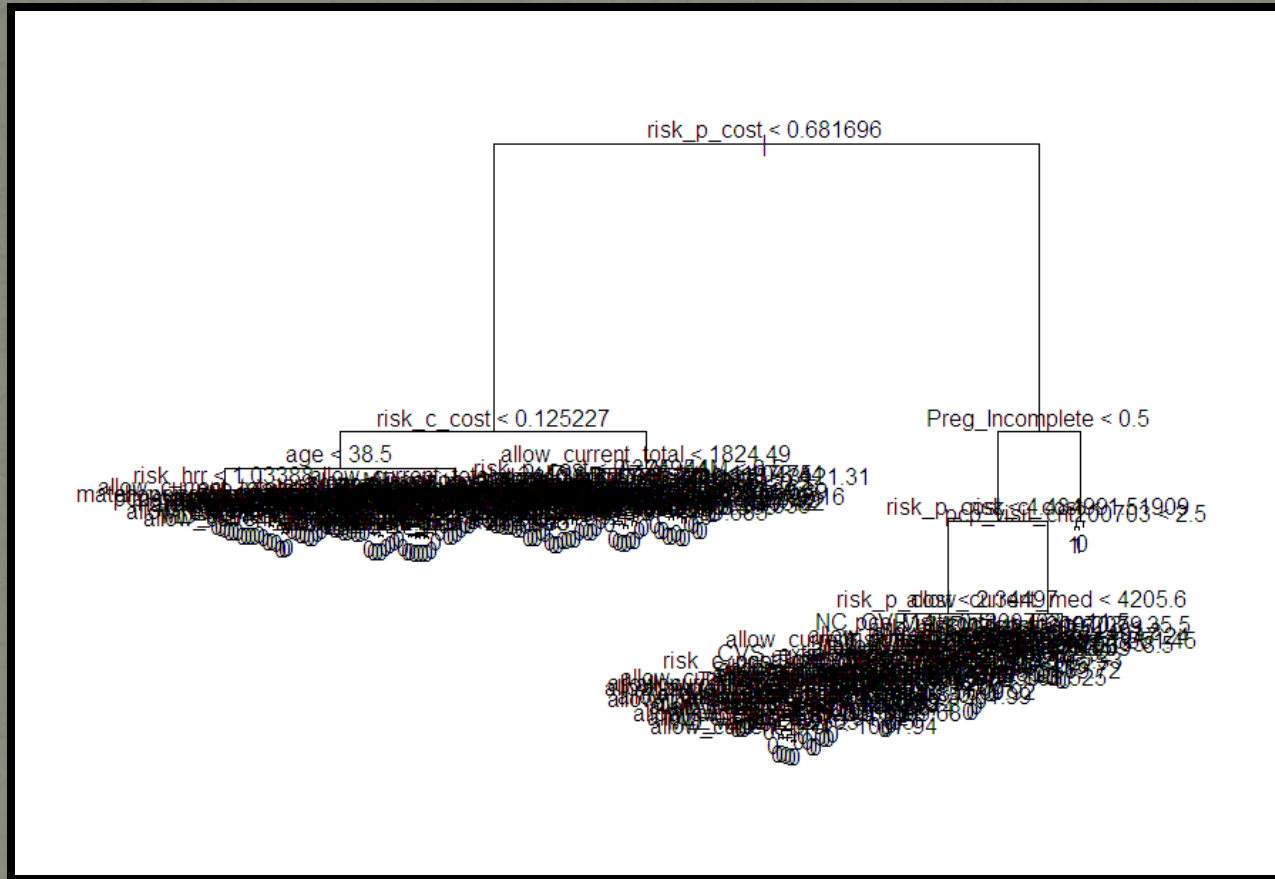
SAS EM: Tree Size Selection



Compare ROC Curves

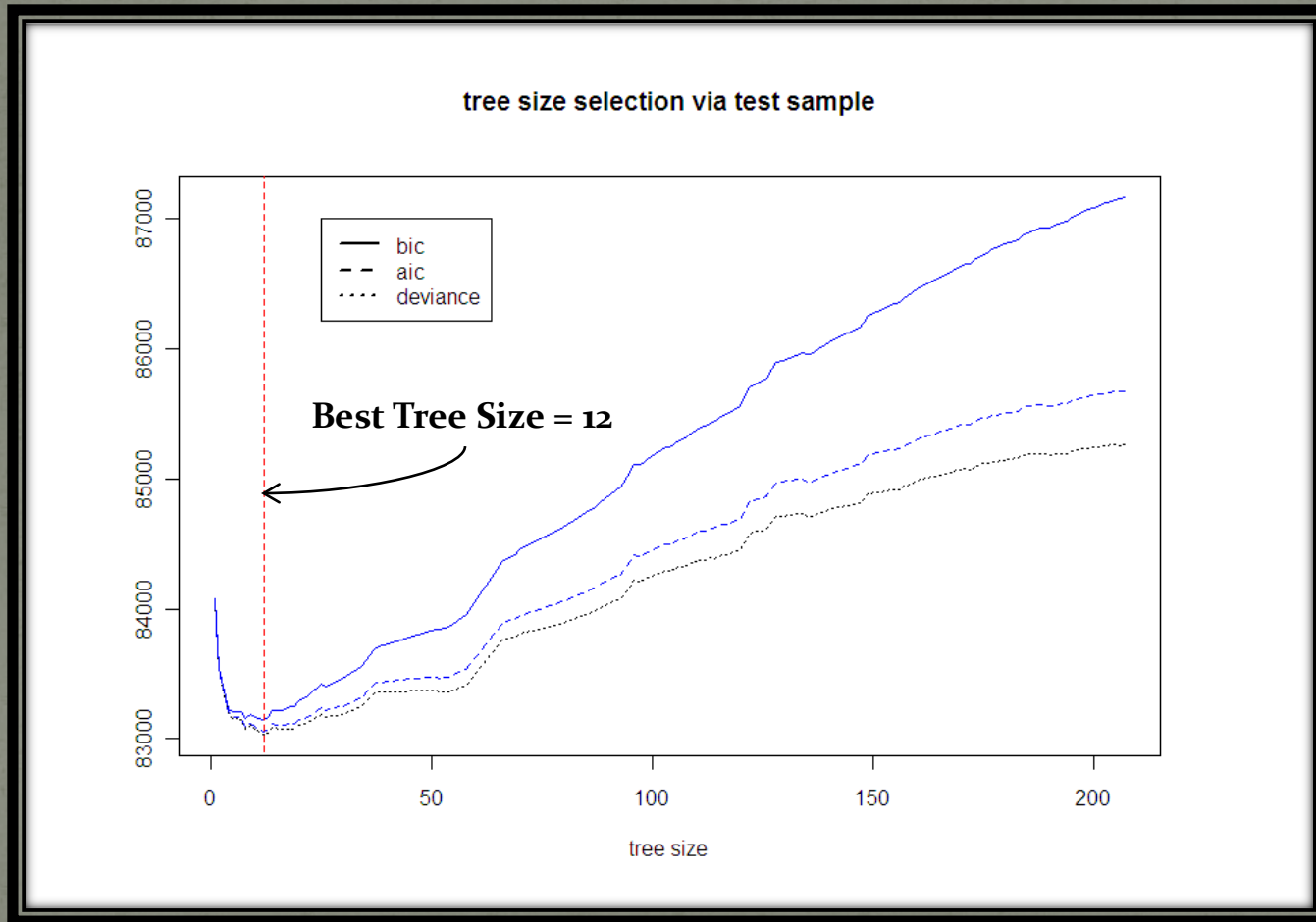


R: Initial Large Tree



Tree size or the Number of terminal nodes is 207

Tree Size Selection via Validation Sample



The Best Tree (Size = 12)

- The Final Tree Structure

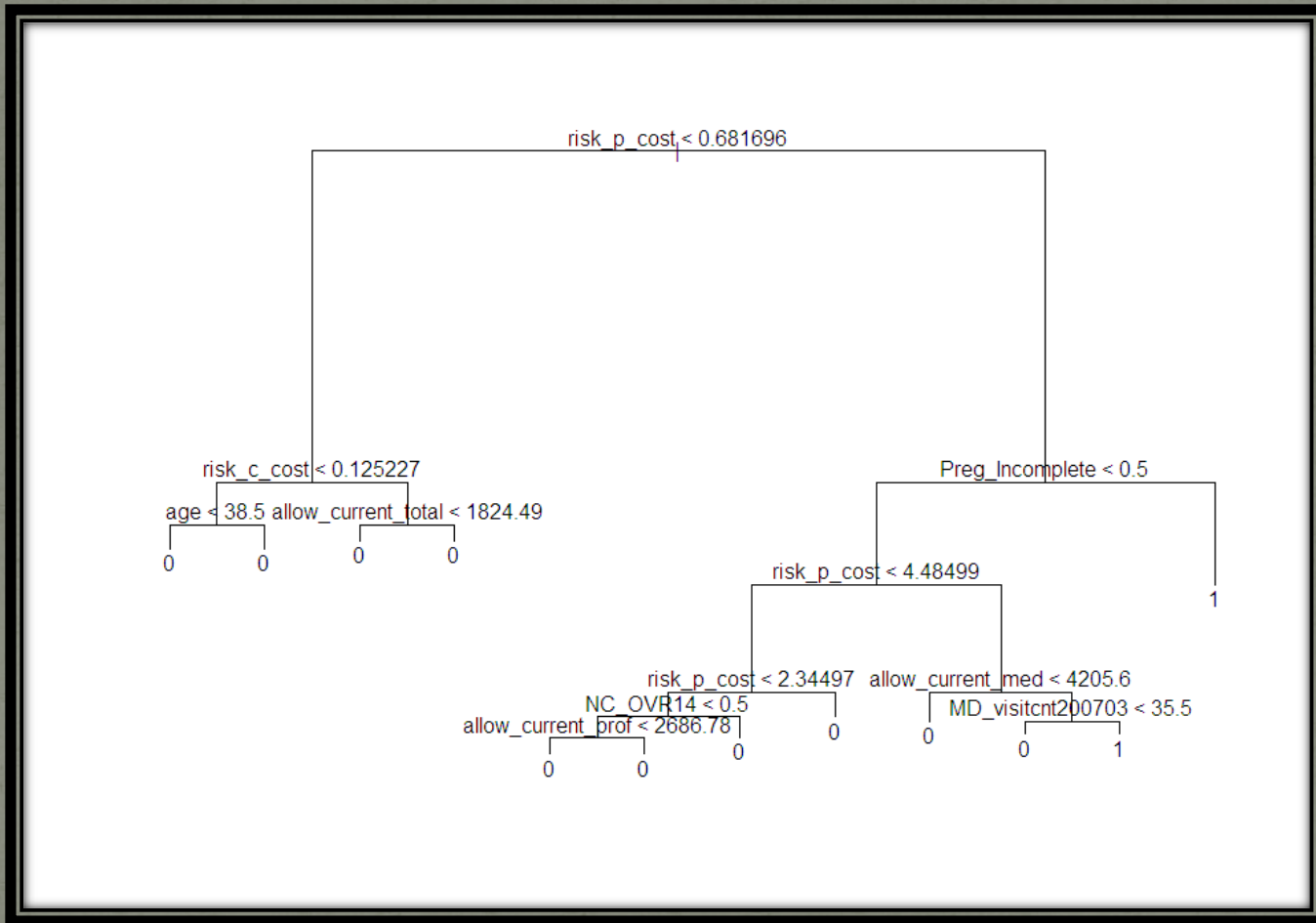
node, split, n, deviance, yval, (yprob) ----- * denotes terminal node

```
1) root 14955 6803.00 0 ( 0.93982 0.06018 )
  2) risk_p_cost < 0.681696 10037 2821.00 0 ( 0.96832 0.03168 )
    4) risk_c_cost < 0.125227 6327 1329.00 0 ( 0.97819 0.02181 )
      8) age < 38.5 2602 300.40 0 ( 0.98962 0.01038 ) *
      9) age > 38.5 3725 998.60 0 ( 0.97020 0.02980 ) *
    5) risk_c_cost > 0.125227 3710 1440.00 0 ( 0.95148 0.04852 )
      10) allow_current_total < 1824.49 2672 872.70 0 ( 0.96145 0.03855 ) *
      11) allow_current_total > 1824.49 1038 548.70 0 ( 0.92582 0.07418 ) *
  3) risk_p_cost > 0.681696 4918 3576.00 0 ( 0.88166 0.11834 )
    6) Preg_Incomplete < 0.5 4783 3264.00 0 ( 0.89254 0.10746 )
      12) risk_p_cost < 4.48499 4238 2497.00 0 ( 0.91340 0.08660 )
        24) risk_p_cost < 2.34497 3069 1578.00 0 ( 0.92864 0.07136 )
          48) NC_OVR14 < 0.5 2959 1438.00 0 ( 0.93410 0.06590 )
            96) allow_current_prof < 2686.78 2260 949.60 0 ( 0.94602 0.05398 ) *
            97) allow_current_prof > 2686.78 699 467.90 0 ( 0.89557 0.10443 ) *
          49) NC_OVR14 > 0.5 110 115.40 0 ( 0.78182 0.21818 ) *
        25) risk_p_cost > 2.34497 1169 888.20 0 ( 0.87340 0.12660 ) *
      13) risk_p_cost > 4.48499 545 635.50 0 ( 0.73028 0.26972 )
        26) allow_current_med < 4205.6 147 92.46 0 ( 0.90476 0.09524 ) *
        27) allow_current_med > 4205.6 398 507.10 0 ( 0.66583 0.33417 )
          54) MD_visitcnt200703 < 35.5 349 425.20 0 ( 0.70201 0.29799 ) *
          55) MD_visitcnt200703 > 35.5 49 66.27 1 ( 0.40816 0.59184 ) *
    7) Preg_Incomplete > 0.5 135 187.10 1 ( 0.49630 0.50370 ) *
```

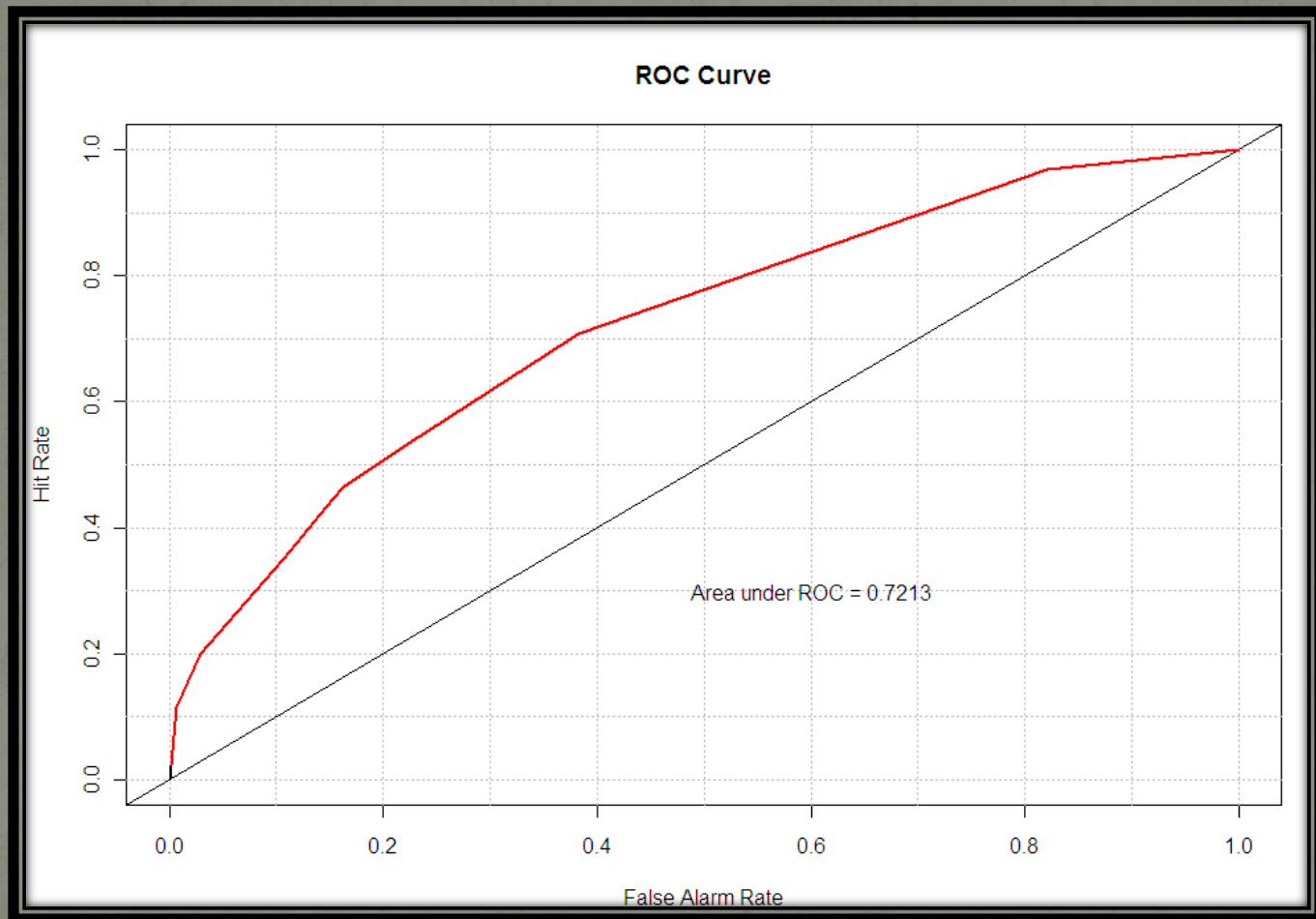
- Variables actually used in tree construction:

```
"risk_p_cost" "risk_c_cost" "age" "allow_current_total" "Preg_Incomplete" "NC_OVR14"
"allow_current_prof" "allow_current_med" "MD_visitcnt200703"
```

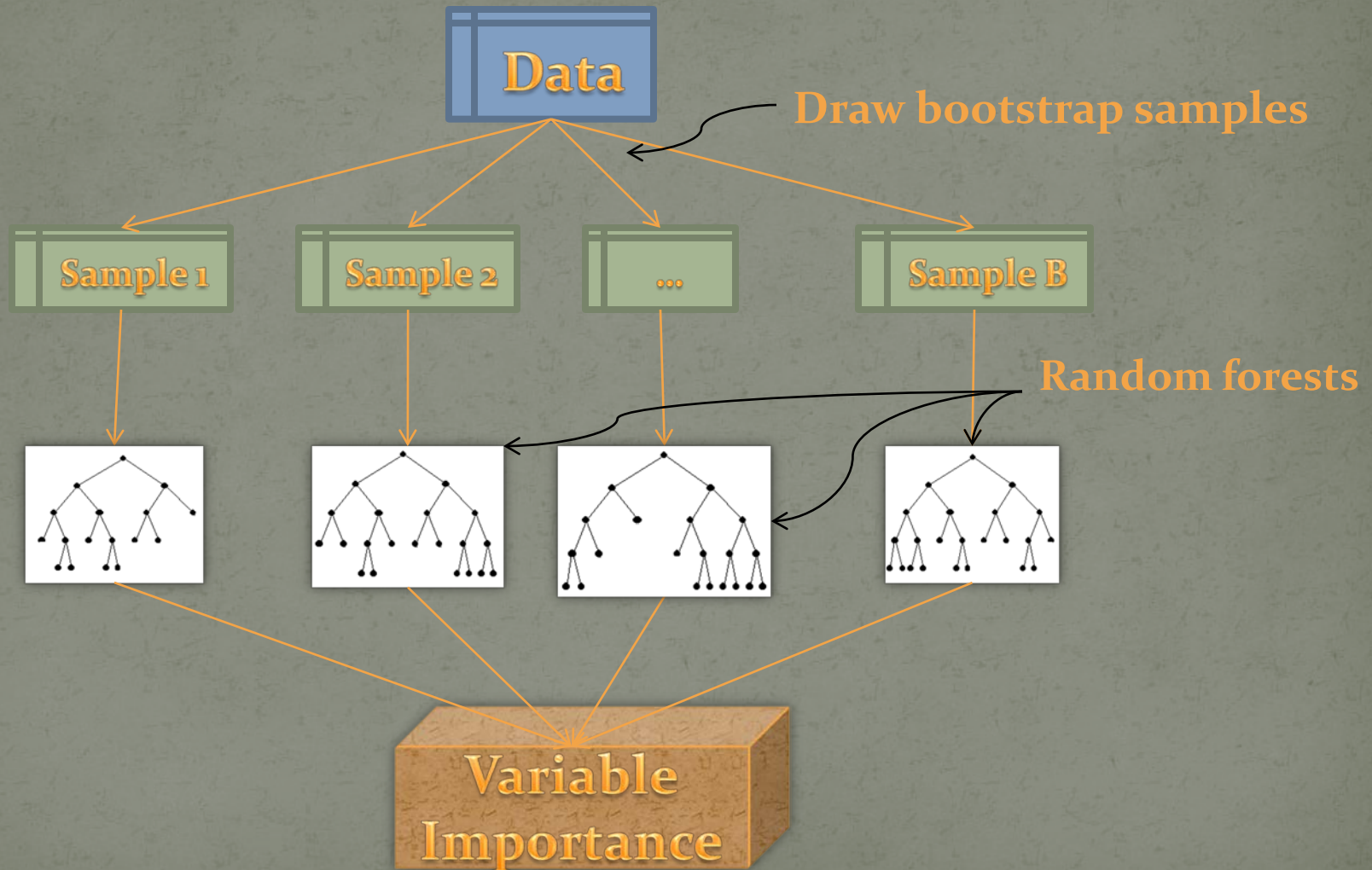
Best Tree



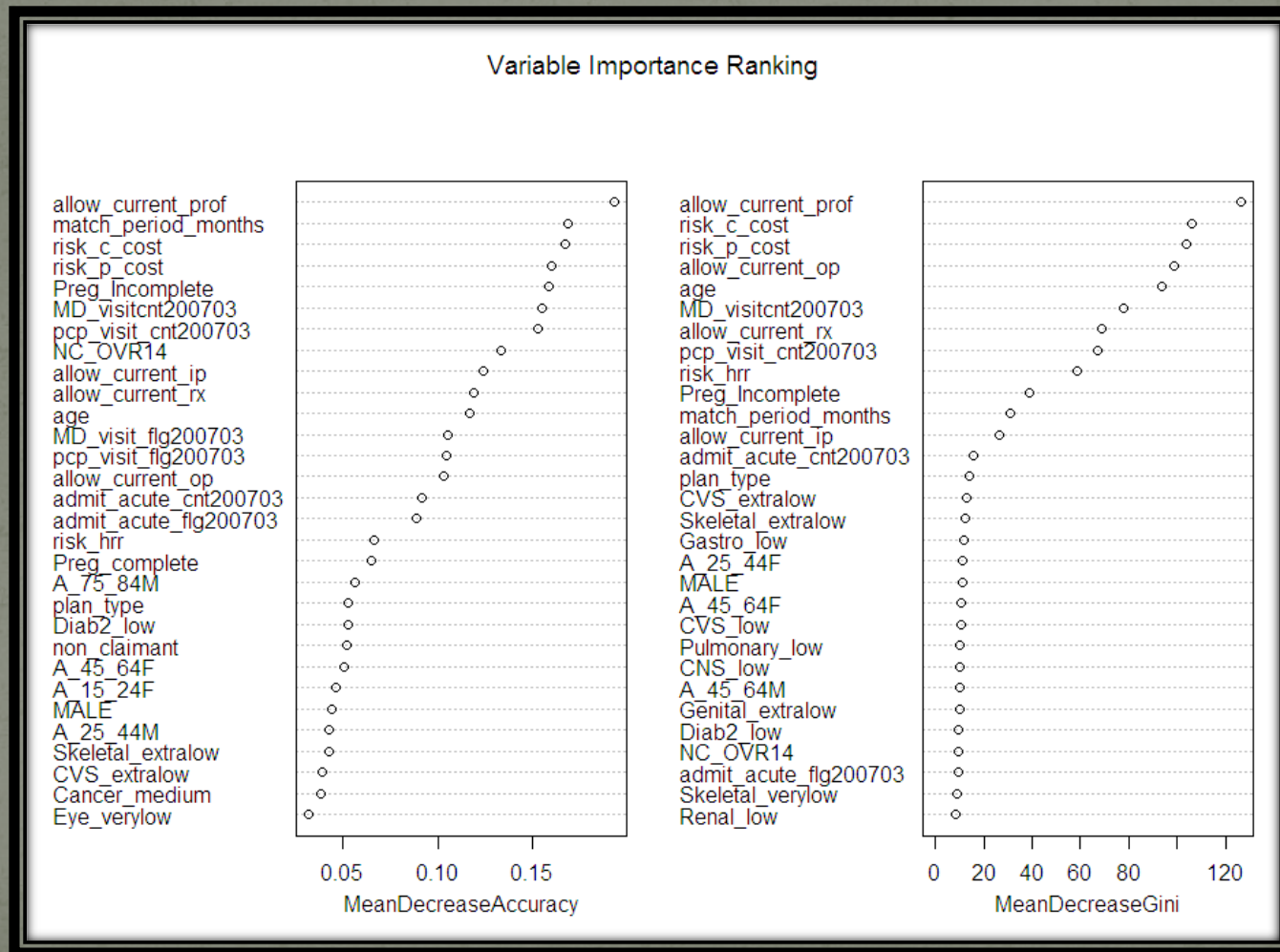
ROC (Test Sample)



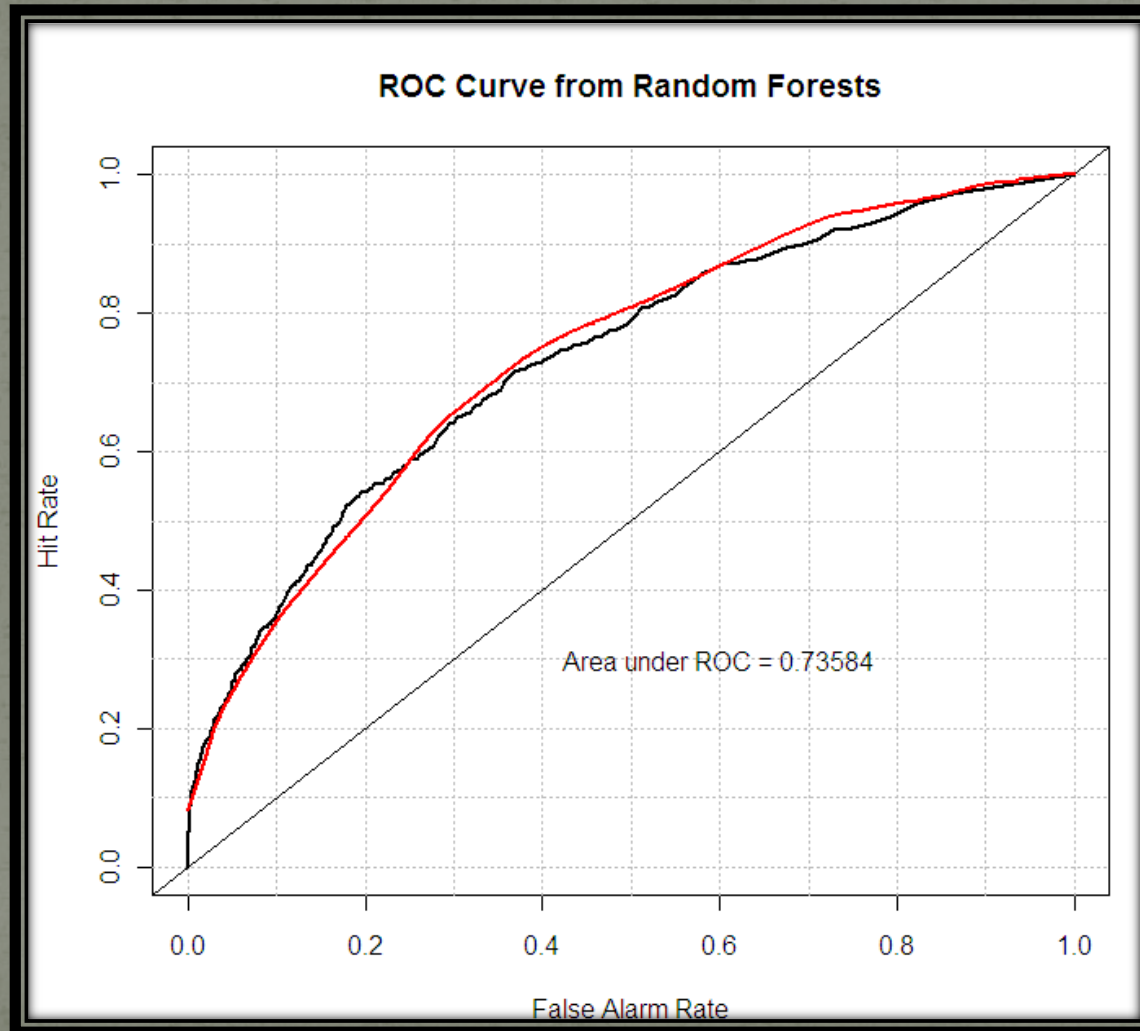
Random Forests



Variable Importance Ranking via Random Forest



The ROC Curve from Random Forests



Thanks!