Random Forest Survival

John Ehrlinger Cleveland Clinic

February 25, 2015

Random Forest

Mature statistical "machine learning" method for

- Regression (continuous outcomes)
- Classification (categorical outcomes)
- Survival (time to event outcomes)
- Others (competing risk, unsupervised, etc.)

Optimized to minimize prediction error

Consistently outperforms other "off the shelf" methods

Random Forest

Ensemble of decision trees

- Democratic method
- Individual weak learners
- Aggregate to a strong learner

Non-parametric

- No model assumptions
- Nonlinear
- Interactions

Data

Data set has:

- n observations
- p independent variables

Ideally, want $n \to$ everyone (unrealistic)

Instead simulate with the Bootstrap

- Randomly select n observations with replacement (b)
- On average 36.8% left out of bootstrap (oob)

Random Forest

Grow a collection of independent decision trees

- One for each Bootstrap data set
- Test with the associated oob data set

But decision trees are

- Inherently unstable
- Tend to over fit training data

They are an ideal weak learner suitable for RF application

Growing a Decision Tree

Recursively partition the data

- Split data nodes (set) into two daughter nodes
- Repeat to exhaustion

Two requirements

- Split rule
- Stopping rule

Growing a Decision Tree

Split rule

Test each variable for optimal node segmenting

- Optimize over classes of categorical variables
- Optimize along values of continuous variables

Choose optimal variable

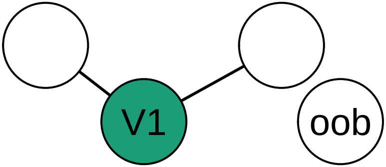
Dependent on the problem domain

- Regression - MSE
- Classification - Gini index (Generalize Binomial Variance)
- Survival - Log-rank

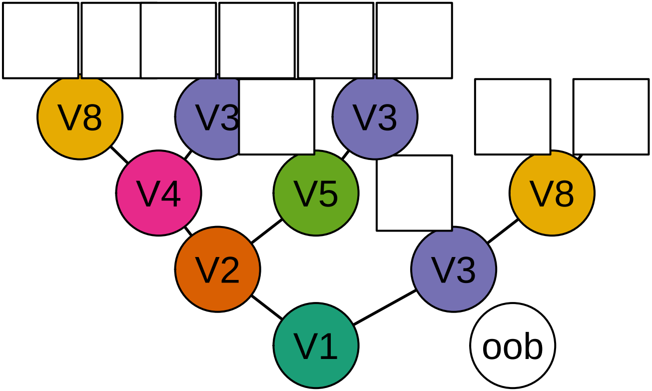
Optimally segregate two groups of observations

Growing a Decision Tree

Growing a Decision Tree



Growing a Decision Tree



Growing a Decision Tree

Stopping Rule defines Terminal Nodes

- Minimal number of members
- Homogeneity

Defaults depend on the problem domain

- Regression - min 5 unique cases
- Classification - homogeneous node (min of 1)
- Survival - min 3 unique cases

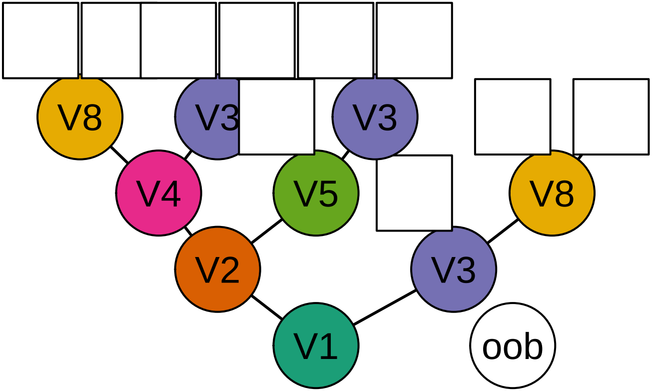
Testing a Decision Tree

Tree sorts each observation into a unique terminal nodes

Test the tree with oob data.

- Sort test observations into terminal nodes
- Predict from training observations
- Compare with test response

Testing a Decision Tree



Decision Tree Prediction

Defined by terminal node membership.

- Fit a model to training set members
- Predict from model

One model for each terminal node within the tree.

Depends on the problem domain

- Regression - mean value
- Classification - probability of class membership
- Survival - Kaplan–Meier estimates

Random Forest Trees

A forest of independent decision trees

- Independent bootstrap training data
- Add extra randomization step

At each node split, RF randomly selects a subset (mtry $\leq p$) of candidate variables for the split rule optimization

Default depends on the problem domain

- Regression - mtry $=$ ceiling($p/3$)
- Classification - mtry $=$ ceiling($\sqrt{p}$)
- Survival - mtry $=$ ceiling($\sqrt{p}$)

Random Forest Prediction

A forest of independent decision trees

- Observations in a terminal node have the same predicted outcome
- Bagging (Bootstrap Aggregation) over all trees

Default depends on the problem domain

- Regression - average estimates
- Classification - voting or average probabilty
- Survival - average survival estimates

Random Forest Performance

Measure of generalization error

- oob data used to calculate forest prediction error

Depends on the problem domain

- Regression - MSE
- Classification - Misclassification error
- Survival - Harrell's concordance index

Breiman's Two Cultures

Machine Learning vs. Statistics

Machine Learning:

- Prediction, Prediction, Prediction
- Black box modeling

Statistics:

- Why?
- Information on underlying process

Random Forest:

- Why not both?
- Insight into the black box of prediction

Random Survival Forest

Extension to time to event data

- Developed at Cleveland Clinic
- Grants and contracts from NHLBI

PBC Example

Primary Biliary Cirrhosis (PBC) of the liver data set (Fleming and Harrington 1991)

Randomized trial of D-penicillamine (DPCA)

Mayo Clinic

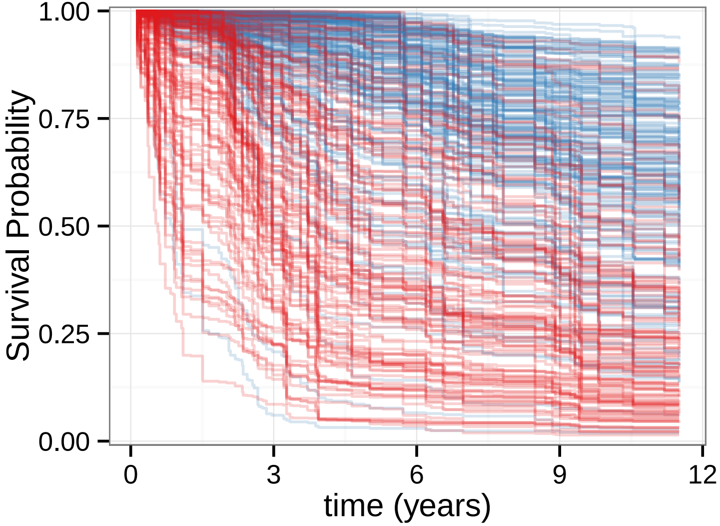
312 patients from 1974 to 1984

- 125 deaths
- 17 variables

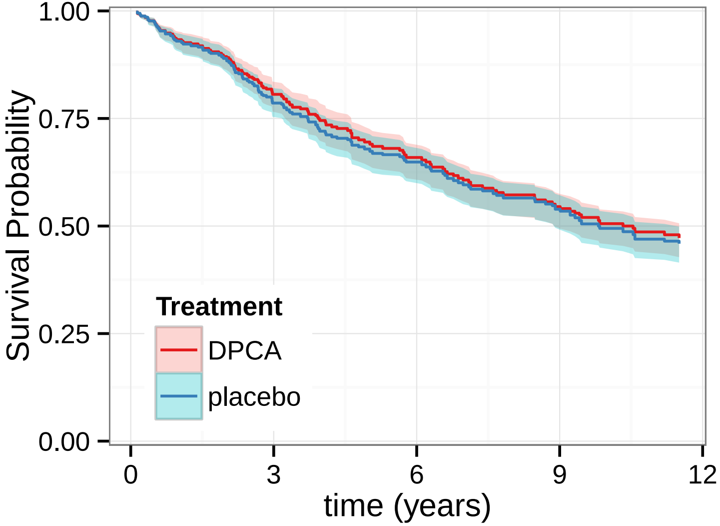
PBC Example

Variables	Coef.	Std. Err.	Z stat.
Age	0.033	0.009	3.84
log(Albumin)	-3.055	0.724	-4.22
log(Bilirubin)	0.879	0.099	8.90
Edema	0.785	0.299	2.62
log(Prothrombin Time)	3.016	1.024	2.95

Random Survival Forest



Random Survival Forest



Variable Selection

Two independent methods

Variable IMPortance (VIMP)

- Based on RF Prediction Error
- Measures the impact of variable misspecification

Minimal Depth

- Property of decision tree construction
- Measures how a variable segments nodes

Variable Selection - VIMP

Prediction error (PE) estimate from oob data

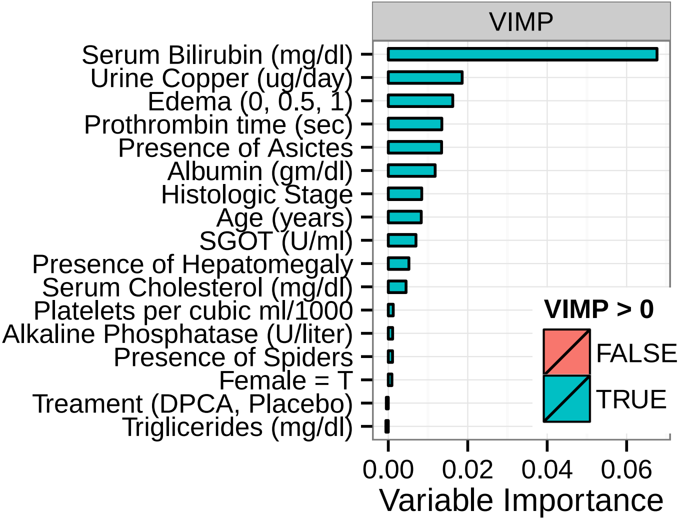
For each variable:

- Randomize values within the variable
- Predict with randomized data
- Calculate a New Prediction Error estimate (NPE)

VIMP = PE - NPE

- Positive value: important in reducing error
- Near zero: no impact on prediction
- Negative value: noise variable

Variable Selection - VIMP

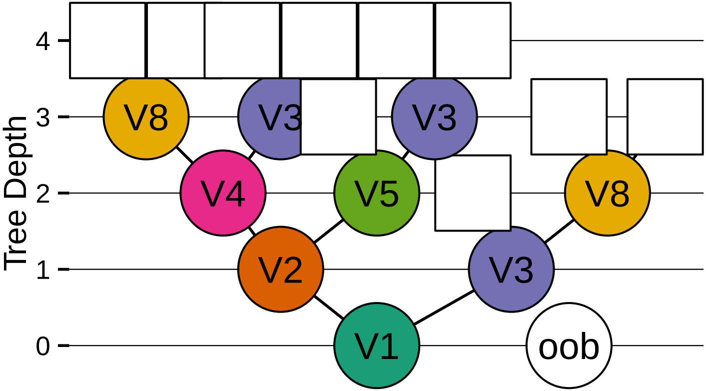


Variable Selection - Minimal Depth

Within each tree

- Number the node split levels
- Find the minimum split level for each variable

Variable Selection - Minimal Depth



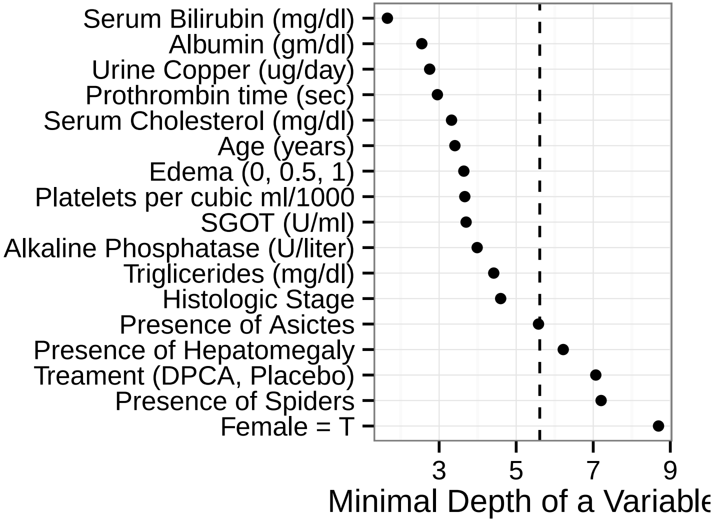
Variable Selection - Minimal Depth

Average minimal split levels

- each variable
- over the forest

Lower values split largest nodes

Variable Selection - Minimal Depth



Random Forest

VIMP and Minimal Depth

- which variables contribute to forest prediction?

Variable dependence

- How does response depend on variables?

Variable Dependence

Two Options:

Variable Dependence

- Observation Based

Partial Dependence

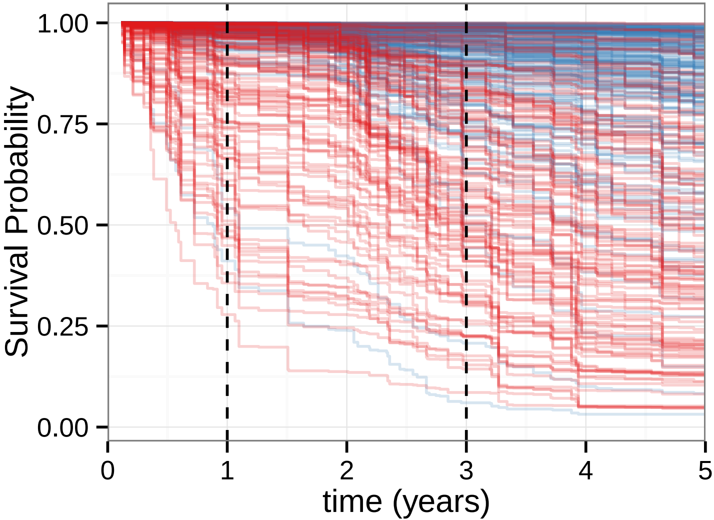
- Population Based

Variable Dependence

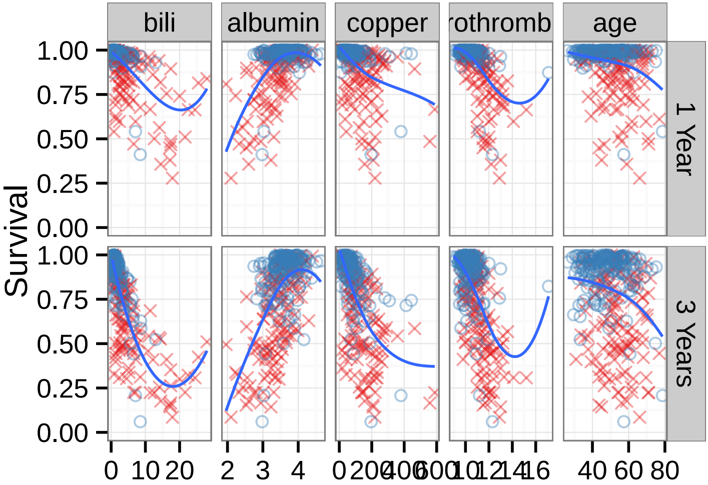
Observation based

- Predicted value for each observation
  - At selected times for survival
- Against variable value

Variable Dependence



Variable Dependence

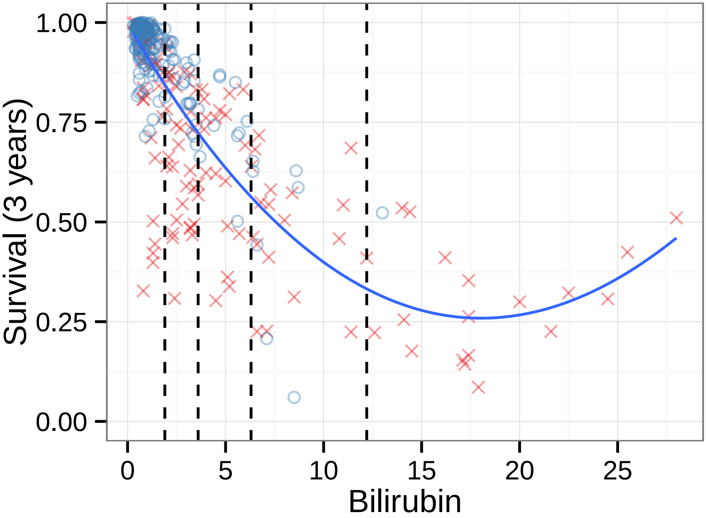


Partial Dependence

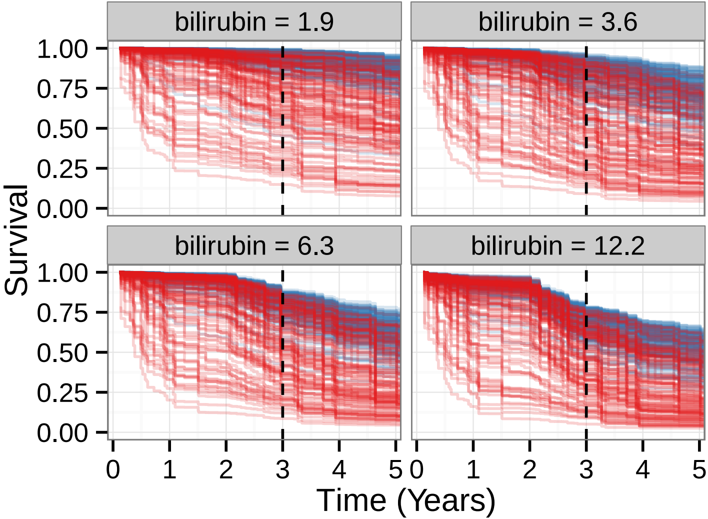
Population Based

- Create nomograms for each observation
  - Across values of variable of interest
  - At selected times for survival
- Average response

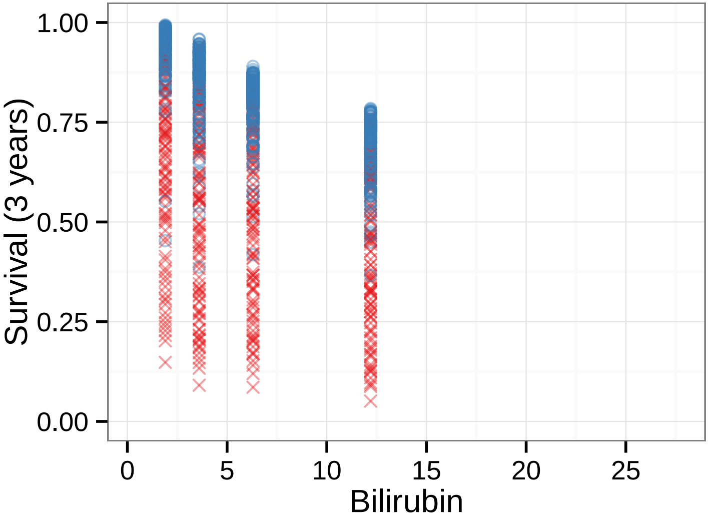
Partial Dependence



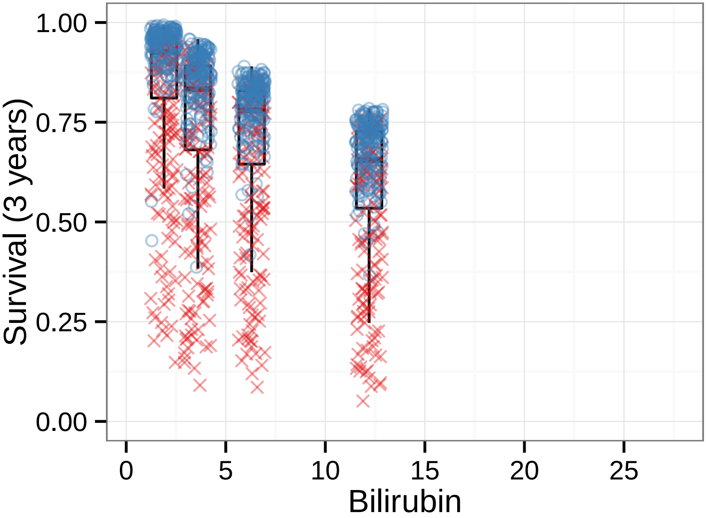
Partial Dependence



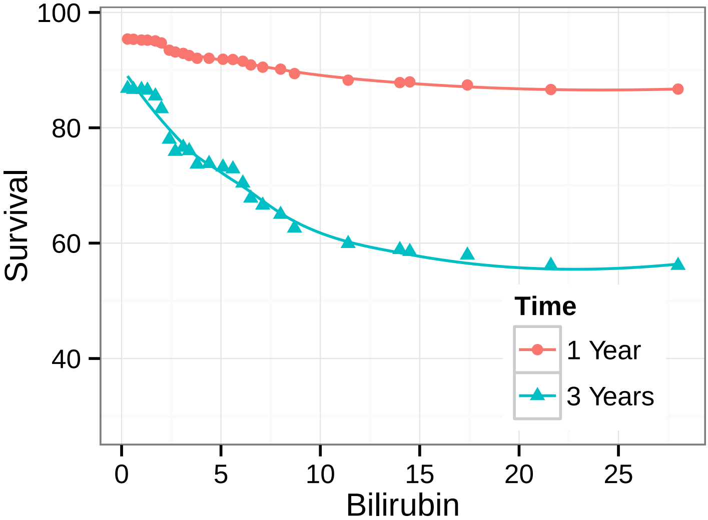
Partial Dependence



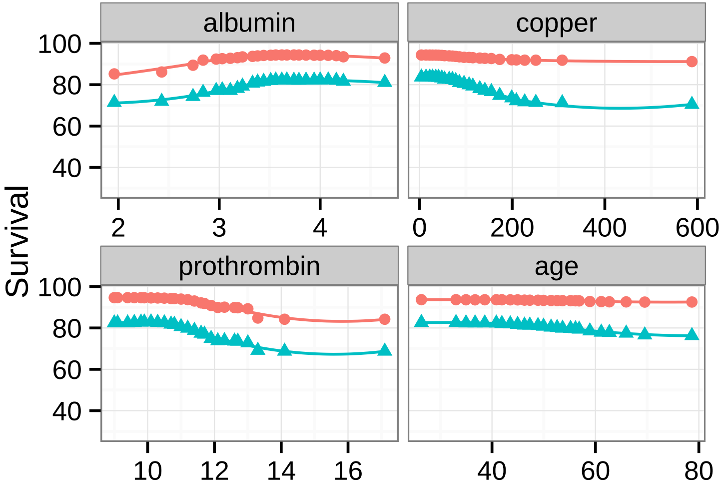
Partial Dependence



Partial Dependence



Partial Dependence



Partial Dependence