



# 3. Regression Trees and Random Forests

Jonathan Hersh (Chapman University Argyros School of Business)

12/7/2020

# Outline

1. Regression Trees
2. Bagging and Ensemble Methods
3. Random Forests

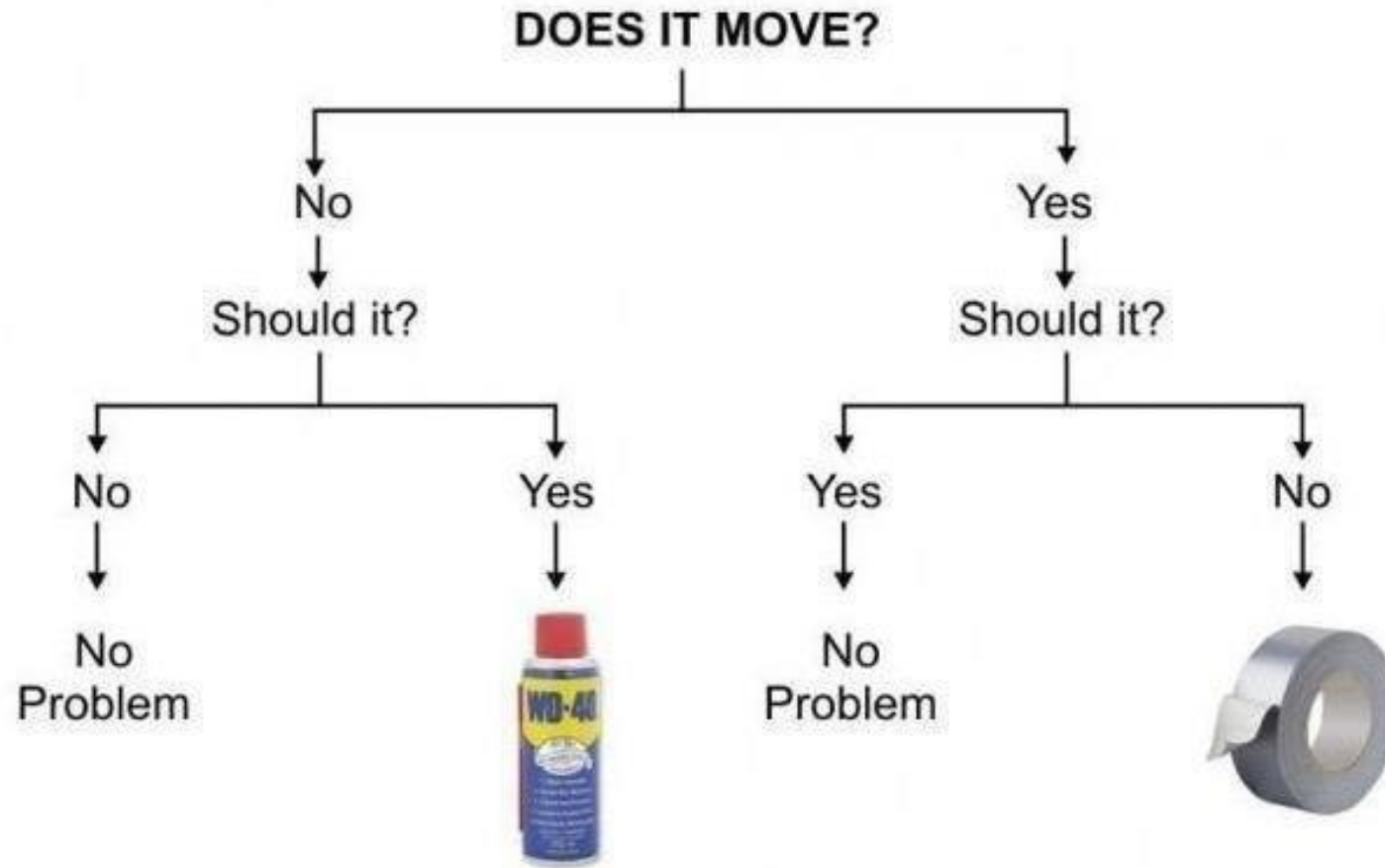


# Regression Trees





# What Are Binary Decision Rules?



- Binary decision rules are any rules with only two options!

# Regression Trees

- Tree based methods *stratify* or *segment* the predictor space into different regions
- Regions are stratified via simple rules
- The splitting rules can be summarized into a tree that is very intuitive



# Pros and Cons of Trees

## Pros

- Simple
- Easy to interpret
- Easy to explain
- Can be displayed graphically!
- Bagging, boosting, and random forests very powerful (combining trees)

## Cons

- Slow with large datasets
- Not easy to use “out of the box”
- Choice of split can be unstable

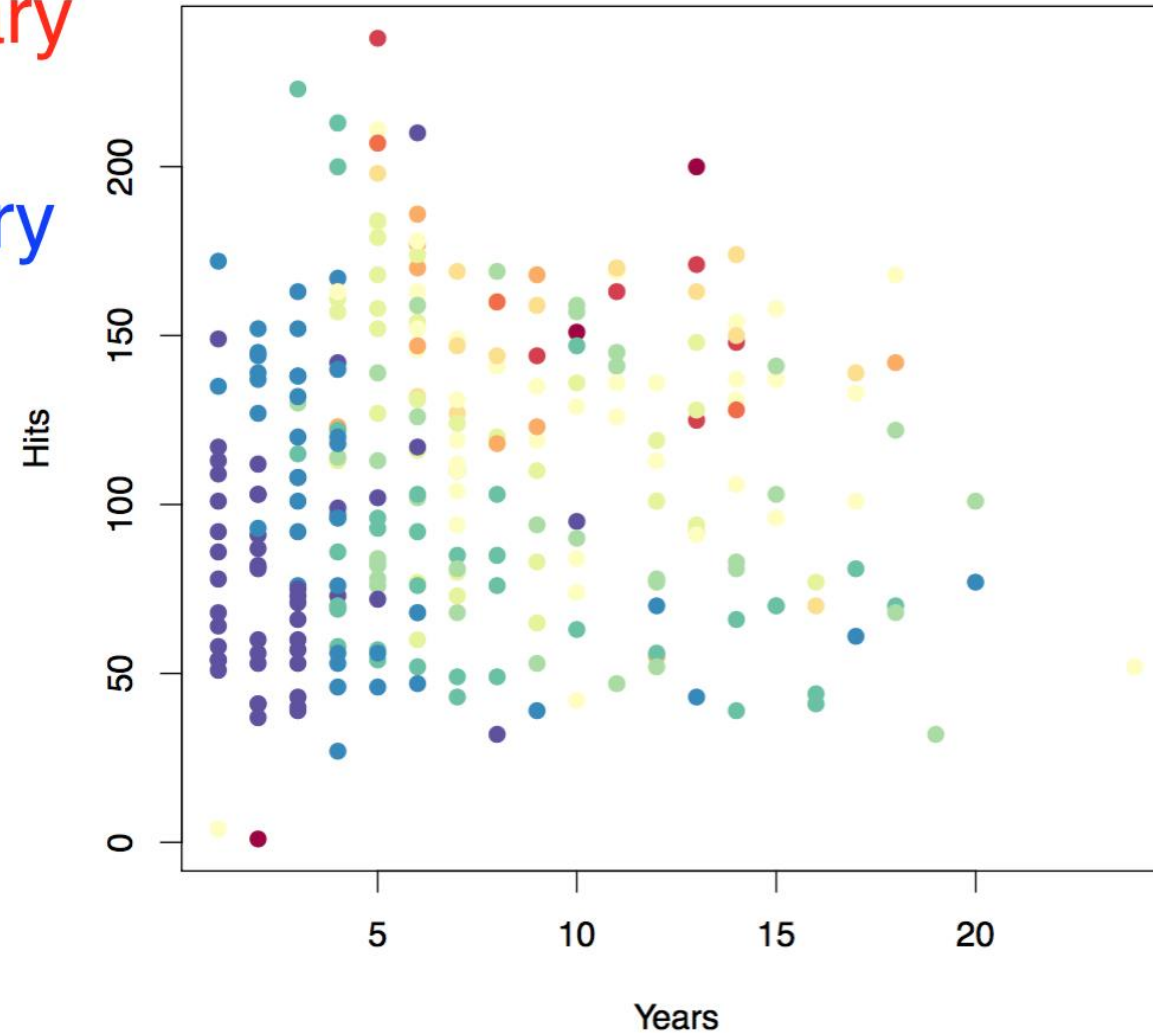
# Decision/Regression Trees

- Decision trees can be applied to both regression problems ( $y_i \in R$ ) and classification problems  $y_i \in \{class1, class2, \dots, \}$
- We'll consider both



# Baseball salary data: how to partition/stratify?

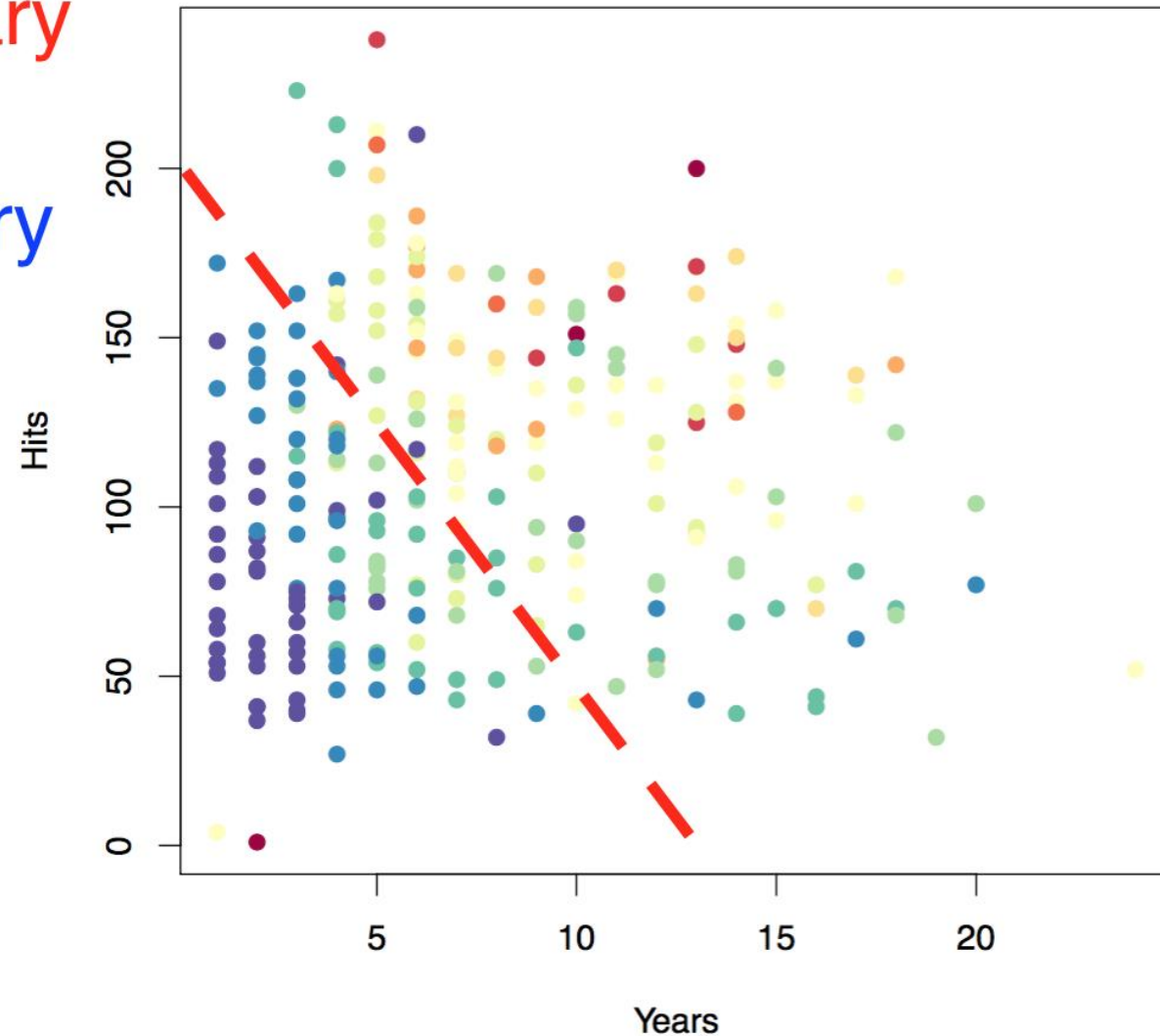
High salary  
red  
Low salary  
blue





# Baseball salary data: how to partition/stratify?

High salary  
red  
Low salary  
blue



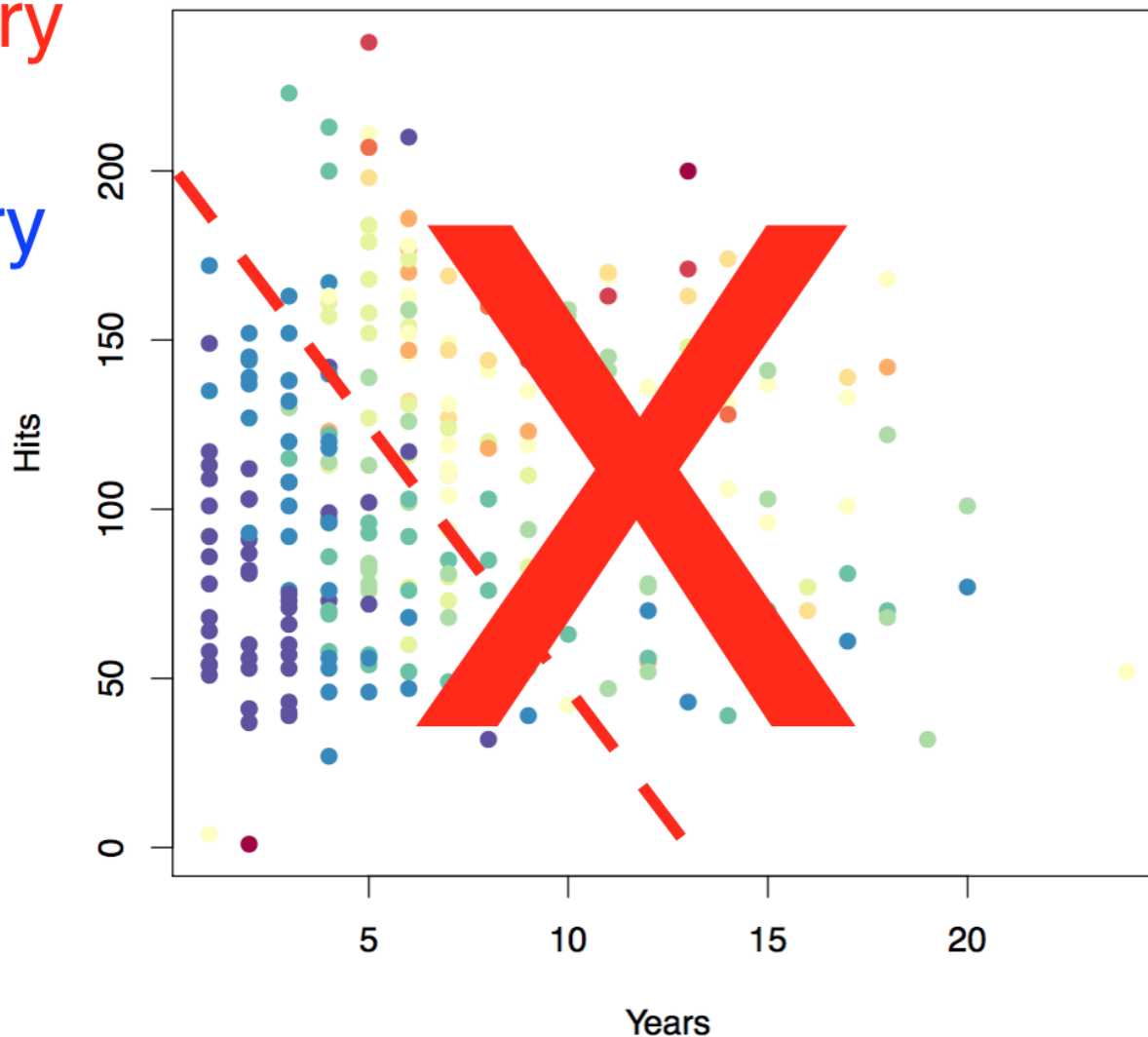
# Baseball salary data: how to partition/stratify?

High salary

red

Low salary

blue



- Only linear classification rules are allowed, e.g.  $\text{year} > 10$

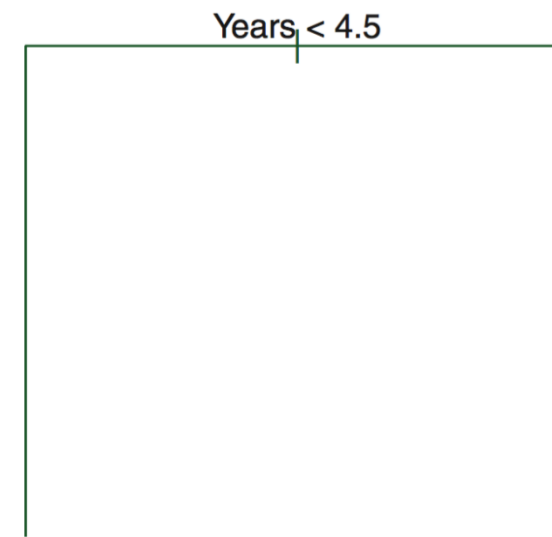
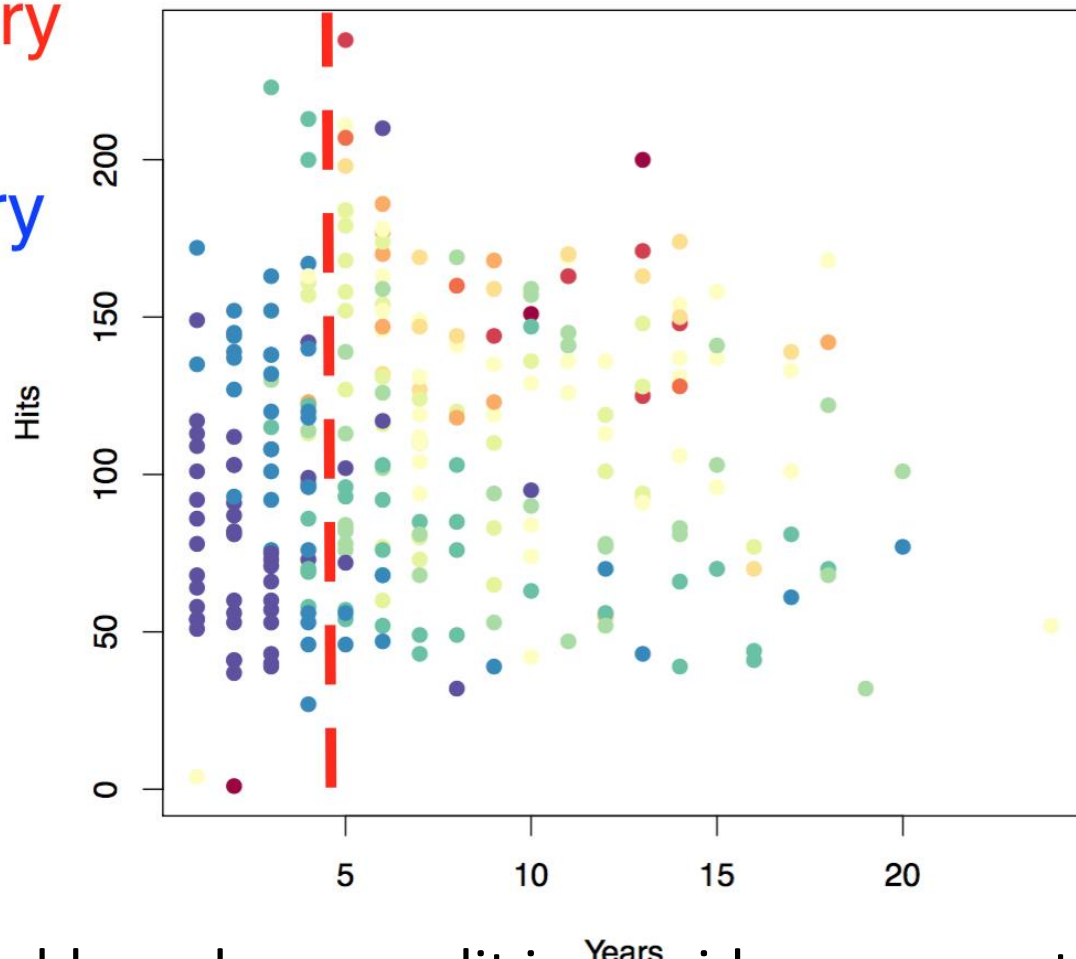
# How Regression Trees are Constructed

High salary

red

Low salary

blue



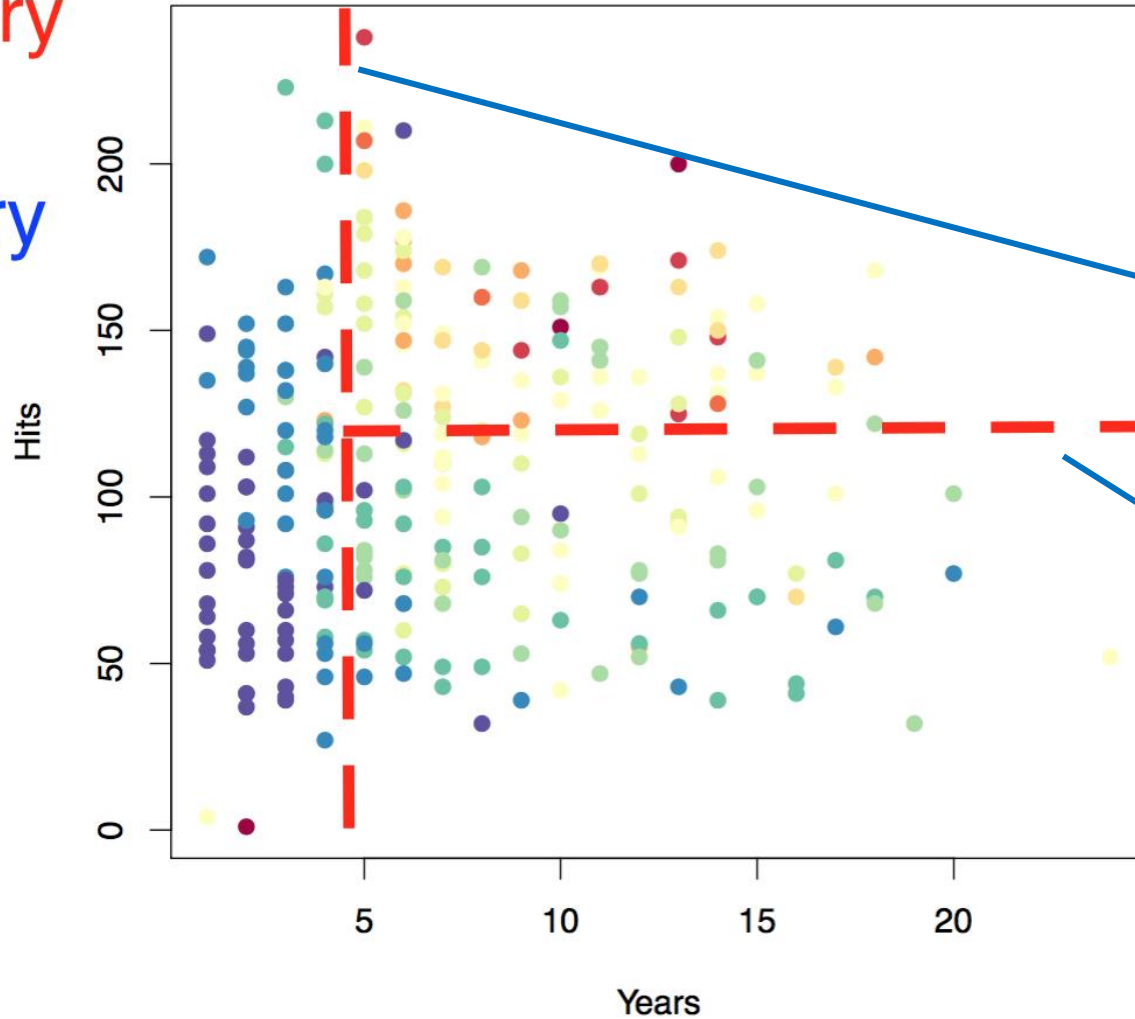
- Every variable and every split is considered.
  - Chosen split is one which maximizes
- separation of high and low salaries:
- Split 1: years > 4.5



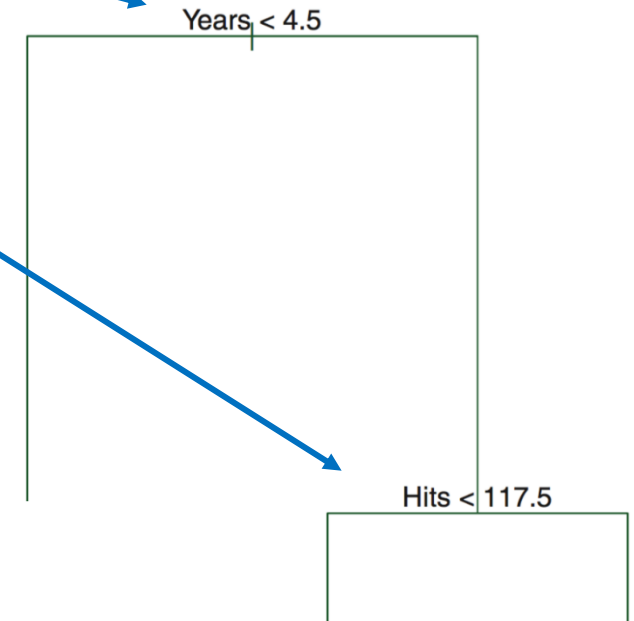
# Baseball salary data: split 2

High salary  
red

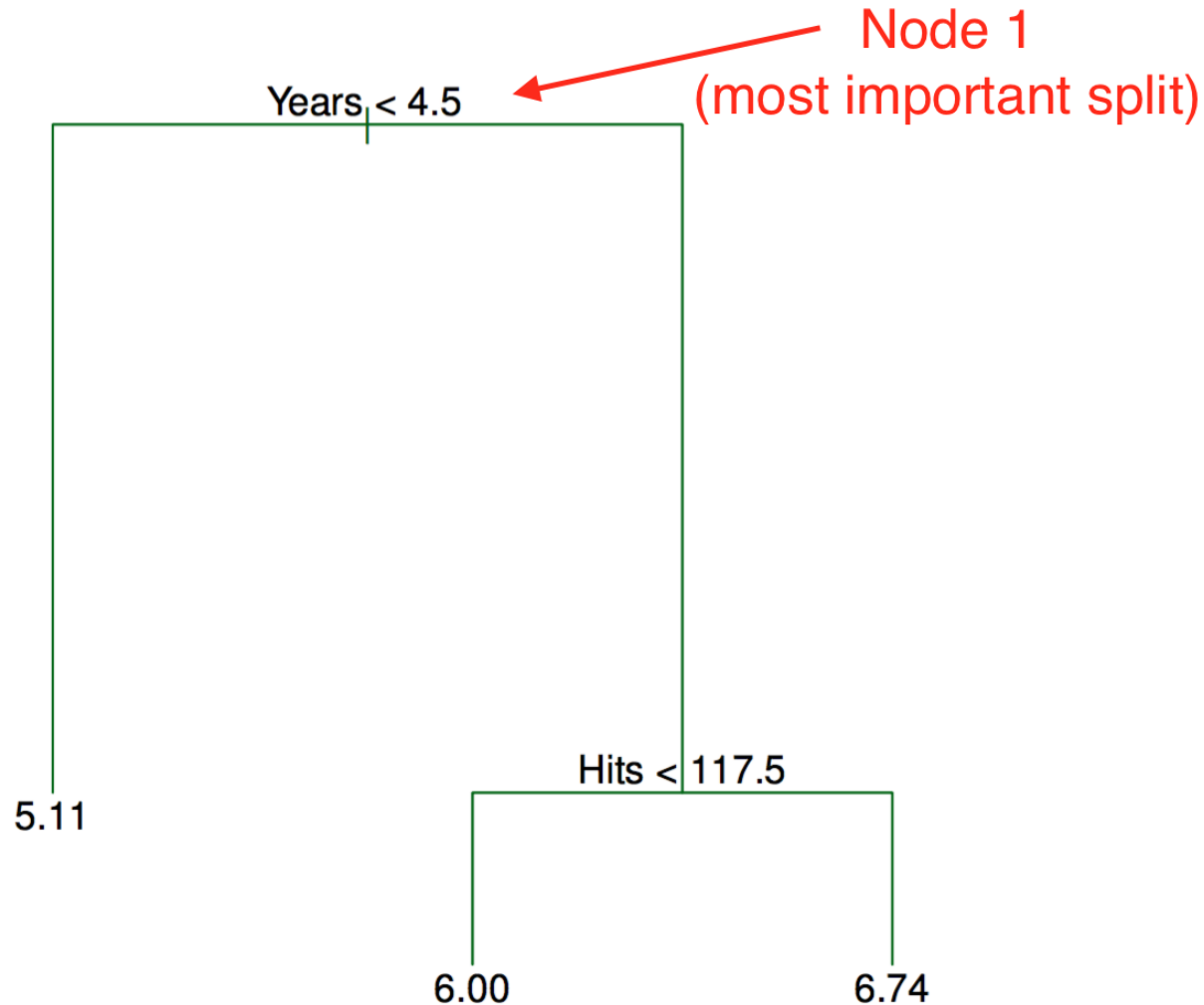
Low salary  
blue



- Split 1: years > 4.5
- Split 2: hits > 117.5

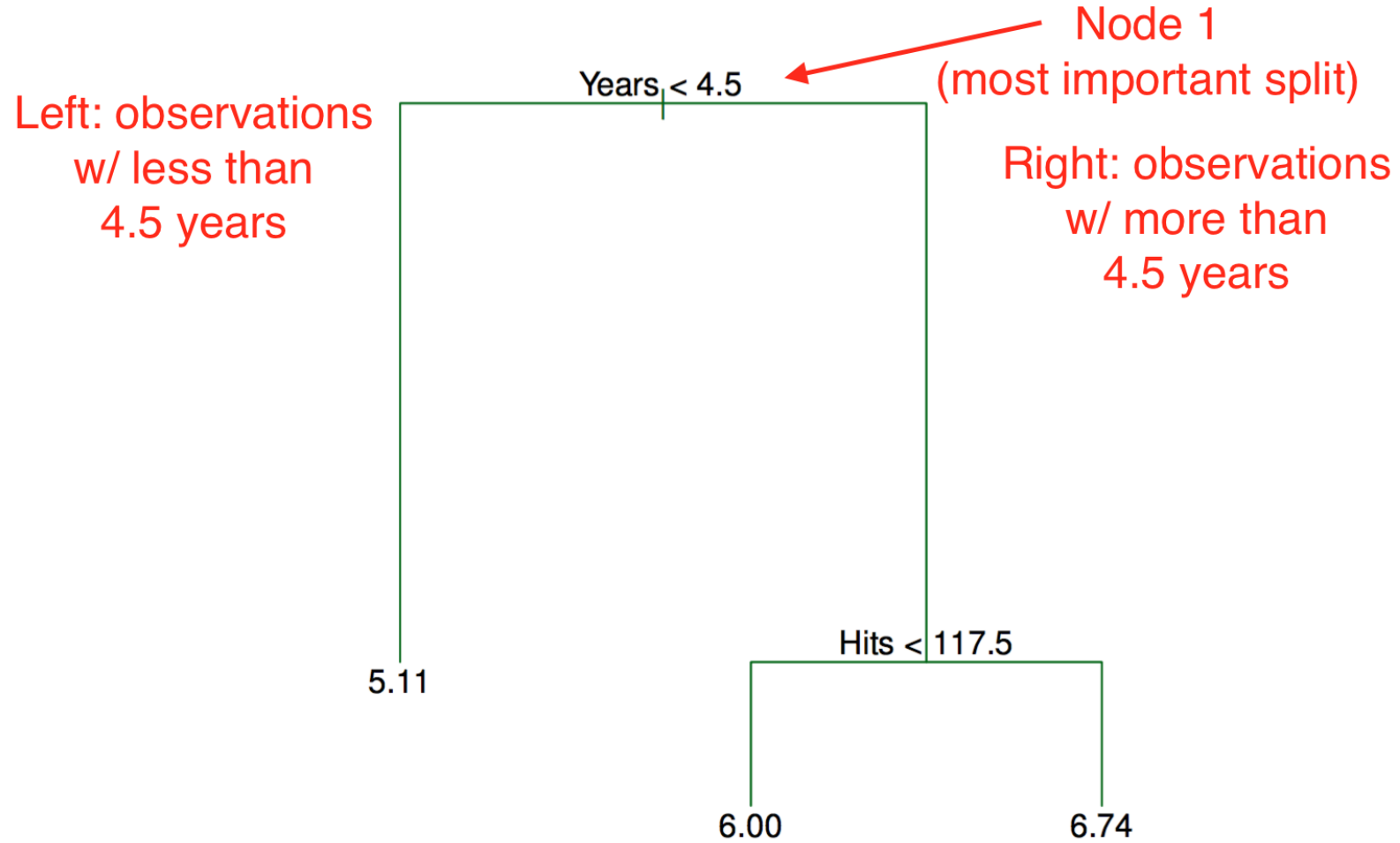


# Tree Representation



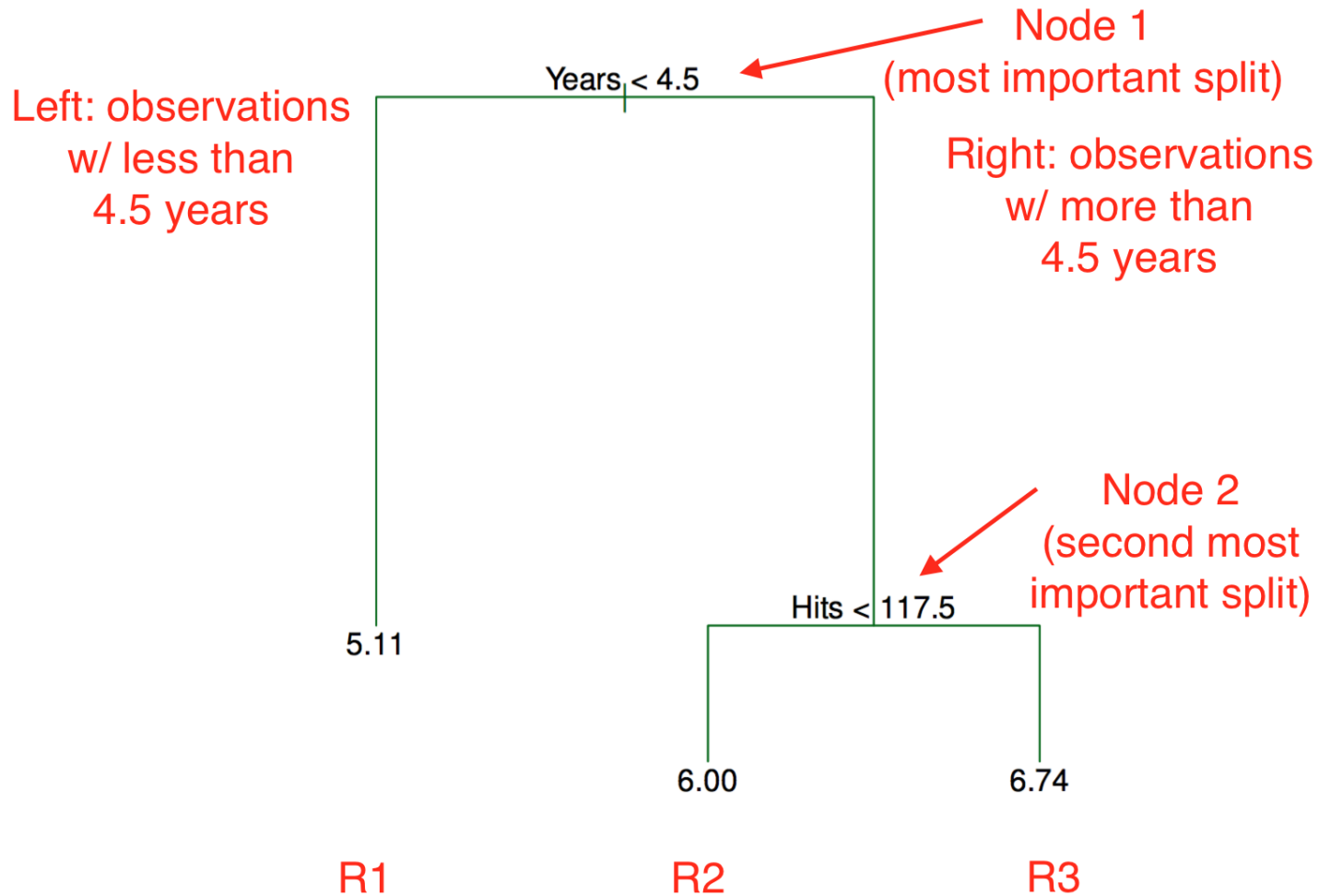
- Trees are read top-down
- Most important split is at top
- Length represents how much within-cluster variance decreases from split
- So Years explains more variance than Hits for this tree

# Tree Representation



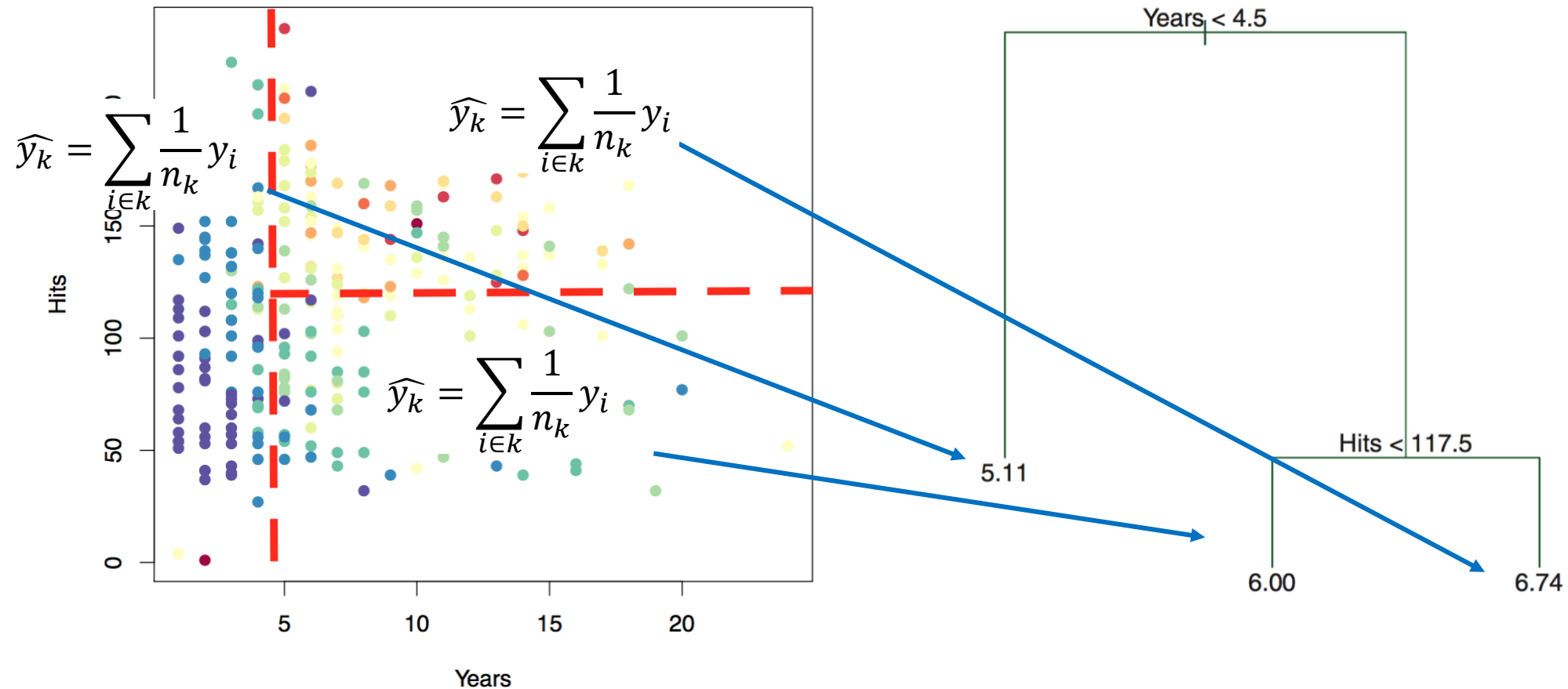


# Predictions for Trees? “Leaf” Values



- At the end of the tree are “leafs” (R1, R2, R3)
- How do we predict using a tree? Using the training data we find the average  $y$  for all the observations in the leaf  $\bar{y}_{leaf} = \frac{1}{n_{leaf}} \sum_{i \in leaf} y_i$
- Any new  $X$ s gets sorted into leafs and assigned the  $y$  average for that leaf:  $\hat{y}_{i \in leaf} = \bar{y}_{leaf}$

# Predictions for Trees? “Leaf” Values



- Leaf predictions are average values in each partition,  $k$

# ctree() function in package “partykit” to build regression tree

ctree {partykit}

R Documentation

## Conditional Inference Trees

### Description

Recursive partitioning for continuous, censored, ordered, nominal and multivariate response variables in a conditional inference framework.

### Usage

```
ctree(formula, data, subset, weights, na.action = na.pass, offset, cluster,  
      control = ctree_control(...), ytrafo = NULL,  
      converged = NULL, scores = NULL, doFit = TRUE, ...)
```

### Arguments

<code>formula</code>	a symbolic description of the model to be fit.
<code>data</code>	a data frame containing the variables in the model.
<code>subset</code>	an optional vector specifying a subset of observations to be used in the fitting process.
<code>weights</code>	an optional vector of weights to be used in the fitting process. Only non-negative integer valued weights are allowed.
<code>offset</code>	an optional vector of offset values.
<code>cluster</code>	an optional factor indicating independent clusters. Highly experimental, use at your own risk.
<code>na.action</code>	a function which indicates what should happen when the data contain missing value.
<code>control</code>	a list with control parameters, see <a href="#">ctree_control</a> .
<code>ytrafo</code>	an optional named list of functions to be applied to the response variable(s) before testing their association with the explanatory variables. Note that this transformation is only performed once for the root node and does not take weights into account. Alternatively, <code>ytrafo</code> can be a function of <code>data</code> and <code>weights</code> . In this case, the transformation is computed for every node with corresponding weights. This feature is experimental and the user interface likely to change.
<code>converged</code>	an optional function for checking user-defined criteria before splits are implemented. This is not to be used and very likely to change.
<code>scores</code>	an optional named list of scores to be attached to ordered factors.
<code>doFit</code>	a logical, if <code>FALSE</code> , the tree is not fitted.
<code>...</code>	arguments passed to <a href="#">ctree_control</a> .



# Estimate a tree model to predict titanic survival

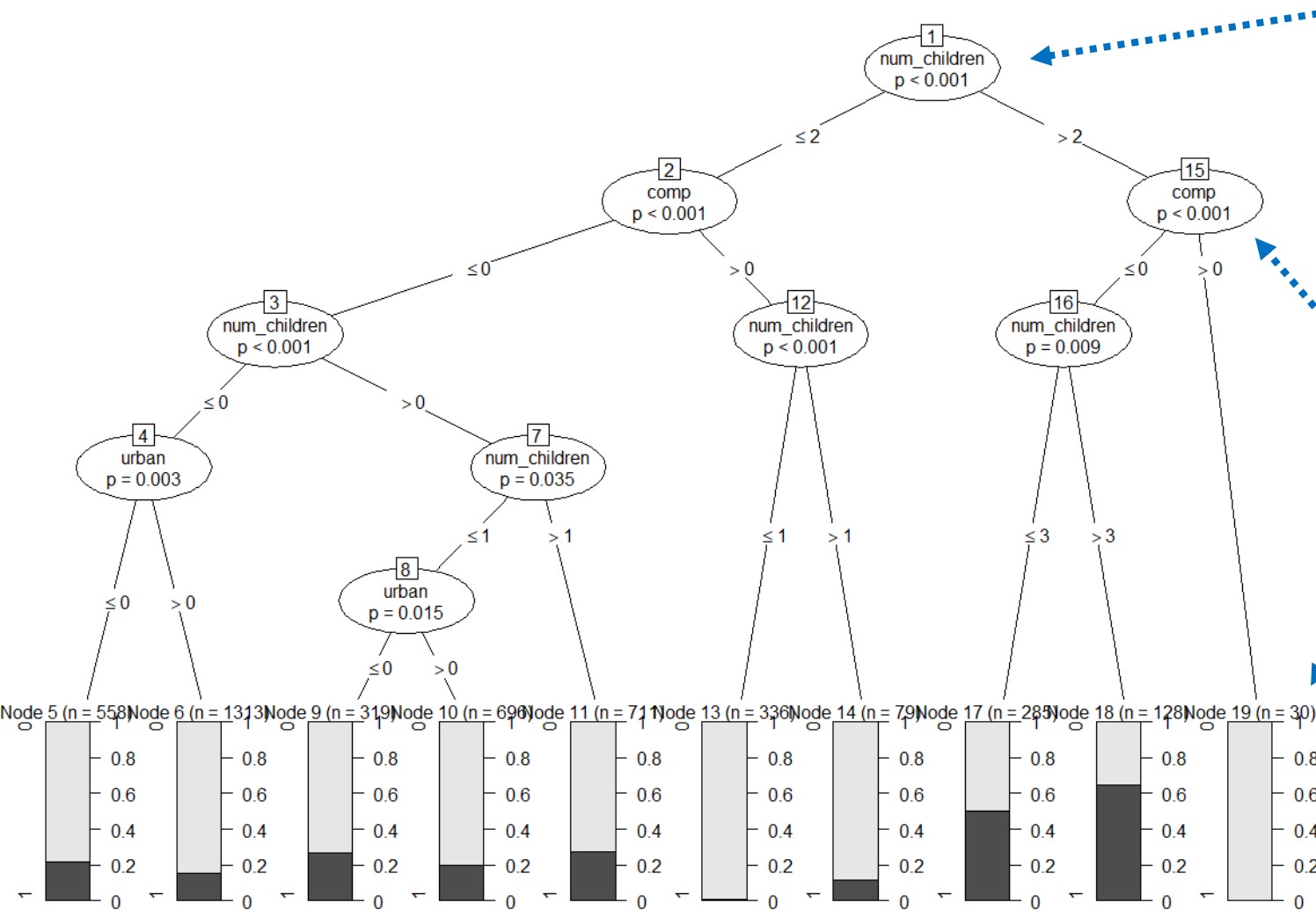
```
# Use the function ctree in rparty to estimate a  
# single regression tree classification model  
poor_tree <- ctree(poor_stat ~ .,  
                   data = CR_train %>%  
                     select(-household_ID))
```

- `ctree()` function estimates a regression tree
- We need: formula (Y and Xs)

```
> print(poor_tree)  
  
Model formula:  
poor_stat ~ num_rooms + bathroom + refrig + no_elect + no_toilet +  
  comp + dep_rate + tv + mobile + num_hh + urban + mean_educ +  
  num_children + num_adults + num_elderly + disabled + mar_stat  
  
Fitted party:  
[1] root  
| [2] dep_rate <= 1.33333  
| | [3] mean_educ <= 7  
| | | [4] num_children <= 0  
| | | | [5] mean_educ <= 5.25: 0.638 (n = 453, err = 104.6)  
| | | | [6] mean_educ > 5.25  
| | | | [7] num_rooms <= 5: 0.789 (n = 341, err = 56.8)  
| | | | [8] num_rooms > 5: 0.949 (n = 78, err = 3.8)  
| | | [9] num_children > 0  
| | | | [10] dep_rate <= 0.625  
| | | | | [11] tv <= 0: 0.663 (n = 193, err = 43.1)  
| | | | | [12] tv > 0: 0.915 (n = 71, err = 5.5)  
| | | | [13] dep_rate > 0.625: 0.509 (n = 344, err = 86.0)  
| | [14] mean_educ > 7  
| | | [15] mean_educ <= 10.33333  
| | | | [16] dep_rate <= 0.33333: 0.916 (n = 521, err = 40.3)  
| | | | [17] dep_rate > 0.33333  
| | | | | [18] comp <= 0: 0.758 (n = 508, err = 93.2)  
| | | | | [19] comp > 0: 1.000 (n = 28, err = 0.0)  
| | | [20] mean_educ > 10.33333  
| | | | [21] num_rooms <= 3: 0.845 (n = 142, err = 18.6)  
| | | | [22] num_rooms > 3  
| | | | | [23] mar_stat in divorced, married: 0.971 (n = 806, err = 22.3)  
| | | | | [24] mar_stat in other, widowed  
| | | | | [25] refrig <= 0: 0.692 (n = 13, err = 2.8)  
| | | | | [26] refrig > 0  
| | | | | | [27] mean_educ <= 12.33333: 0.849 (n = 186, err = 23.8)  
| | | | | | [28] mean_educ > 12.33333  
| | | | | | | [29] dep_rate <= 0.66667: 0.995 (n = 217, err = 1.0)  
| | | | | | | [30] dep_rate > 0.66667: 0.895 (n = 76, err = 7.2)  
| [31] dep_rate > 1.33333  
| | [32] mean_educ <= 12  
| | | [33] dep_rate <= 2.5
```

- Raw model output isn't the prettiest to read (but does show fitted model.)

# Plot Fitted Regression Tree



- First split is shown at top with p-value with null hypothesis that split doesn't increase class "separation". Number of children is the most important variable, and p-value shows it improves model fit
- Second split is given by successive node. Computer access is the second most important
- For each "leaf" at the bottom the distribution plots for the outcome

# Pruning trees

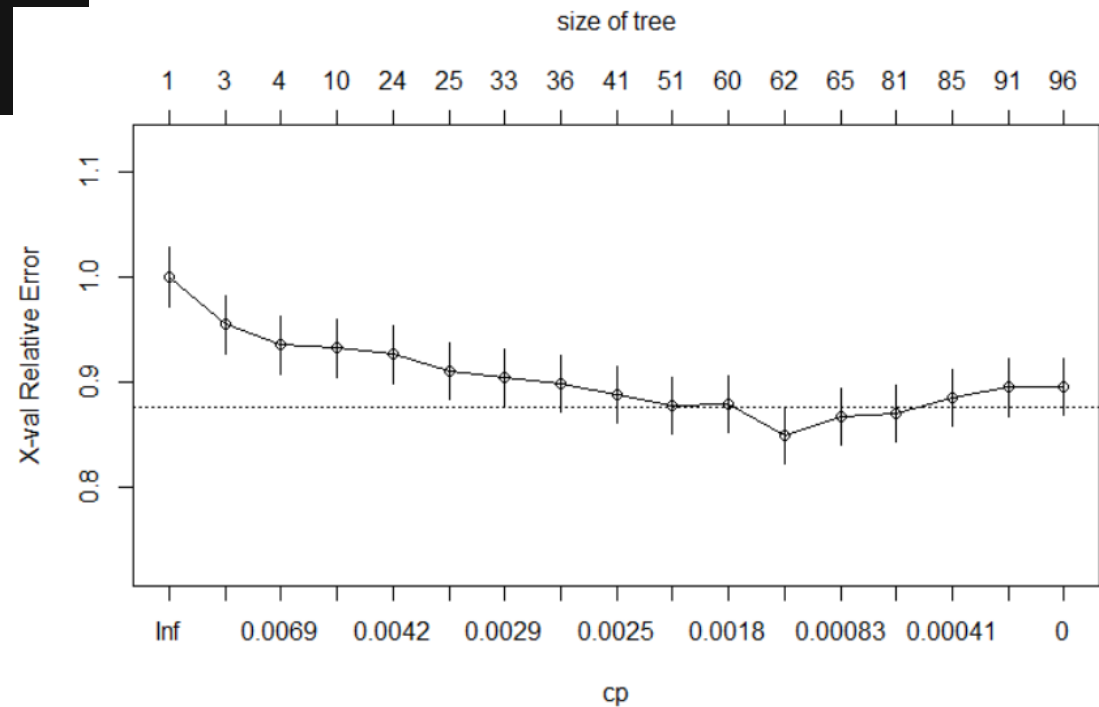
- How do we know when to stop splitting the data?
- **Trees with many splits can overfit the data**
- Solution is to grow a large tree  $T_0$ , then prune it to obtain a smaller sub-tree





# Cross –Validate Our Titanic Data to Determine Optimal Tree Depth

```
poor_rpart <- rpart(poor_stat ~ .,  
  data = CR_train %>%  
  select(-household_ID) %>%  
  mutate(poor_stat = as.factor(poor_stat)),  
  method = "class",  
  control = list(cp = 0,  
    minsplit = 10,  
    maxdepth = 10))  
poor_rpart$cptable
```





# Bagging





## The Netflix Prize: How a \$1 Million Contest Changed Binge-Watching Forever



By DAN JACKSON  
Published On 07/07/2017  
@danielvjackson



In October 2006, Netflix, then a service peddling discs of every movie and TV show under the sun, announced "The Netflix Prize," a competition that lured Mackey and his contemporaries for the computer programmer equivalent of the *Cannonball Run*. The mission: Make the company's recommendation engine 10% more accurate -- or die coding. Word of the competition immediately spread like a virus through comp-sci circles, tech blogs, research communities, and even the mainstream media. ("And if You Liked the Movie, a Netflix Contest May Reward You Handsomely" read the *New York Times* headline.) And while a million dollars created attention, it was the data set -- over 100 million ratings of 17,770 movies from 480,189 customers -- that had number-crunching nuts salivating. There was nothing like it at the time. There hasn't been anything quite like it since.



<https://www.thrillist.com/entertainment/nation/the-netflix-prize>

# Netflix prize

# Netflix Prize

[Home](#)
[Rules](#)
[Leaderboard](#)
[Register](#)
[Update](#)
[Submit](#)
[Download](#)

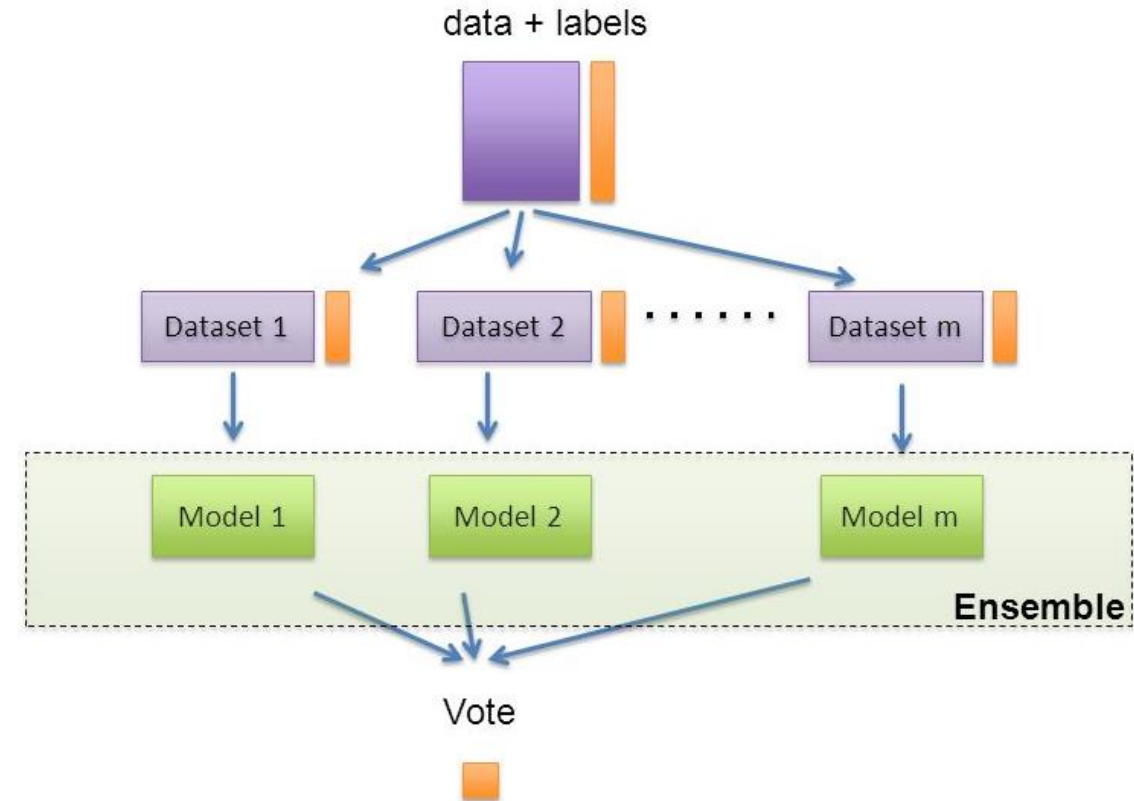
## Leaderboard

10.05%

Display top
 
 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	<a href="#">BellKor's Pragmatic Chaos</a>	0.8558	10.05	2009-06-26 18:42:37
Grand Prize - RMSE <= 0.8563				
2	<a href="#">PragmaticTheory</a>	0.8582	9.80	2009-06-25 22:15:51
3	<a href="#">BellKor in BigChaos</a>	0.8590	9.71	2009-05-13 08:14:09
4	<a href="#">Grand Prize Team</a>	0.8593	9.68	2009-06-12 08:20:24
5	<a href="#">Dace</a>	0.8604	9.56	2009-04-22 05:57:03
6	<a href="#">BigChaos</a>	0.8613	9.47	2009-06-23 23:06:52

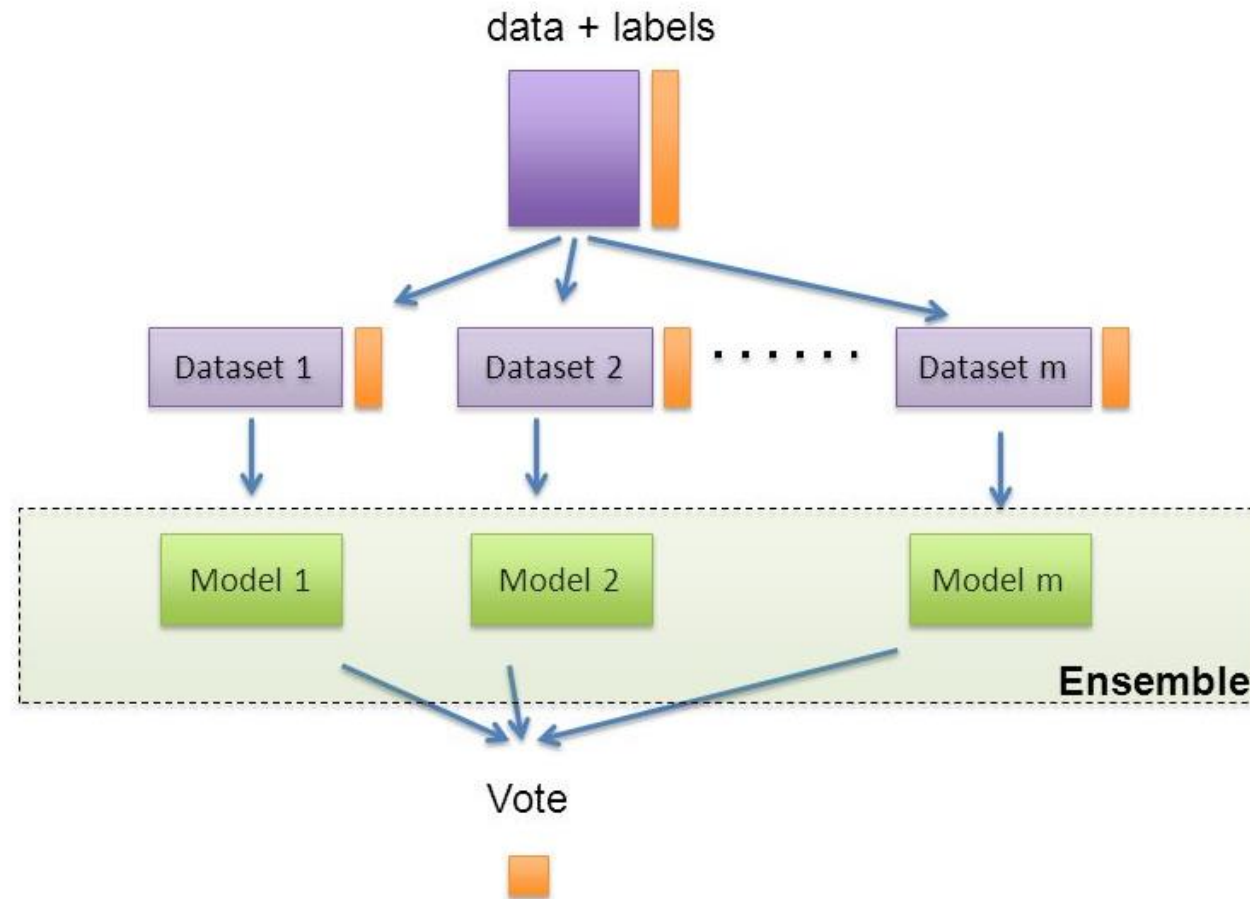
# Netflix prize winners



- Punchline: simple models beat out one very deep model



# Netflix prize conclusion: ensemble of simple methods beats one complex method



# Bagging

- Bagging is short for bootstrap aggregation
- **It's a general purpose method for reducing variance in any machine learning method**
- With  $n$  independent observations,  $z_1, z_2, \dots, z_n$  each with variance  $\sigma^2$ , the variance of the mean ( $\bar{z}$ ) is given by  $\sigma^2/n$
- We usually cannot do this because we don't have multiple training datasets

# Bagging (Bootstrap Aggregation) Algorithm

1. Generate  $B$  bootstrap training datasets
2. Train method on the  $b$ -th bootstrapped data set to obtain  $\hat{f}^*(x)$  the prediction model built using bootstrap sample  $b$
3. Repeat for every bootstrap sample, resulting in  $B$  models, and  $B$  sets of predictions
4. Once all bootstrap samples have models, average all predictions to obtain average prediction over the bootstrapped samples

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*\{b\}}(x)$$

- Netflix prize intuition: average of many models will perform better than a single model

# Out-of-bag error

- Recall that for each bootstrapped sample  $b$  is composed of a subset of the total training data
- For each sample, the data not used to fit the model is referred to as **out-of-bag (OOB) observations**
- We can better approximate out of sample error by only using out-of-bag observations for model validation

preds_boot1	preds_boot2	preds_boot3
0	0	0
0	0	0
NA	0	1
0	NA	0
NA	1	0





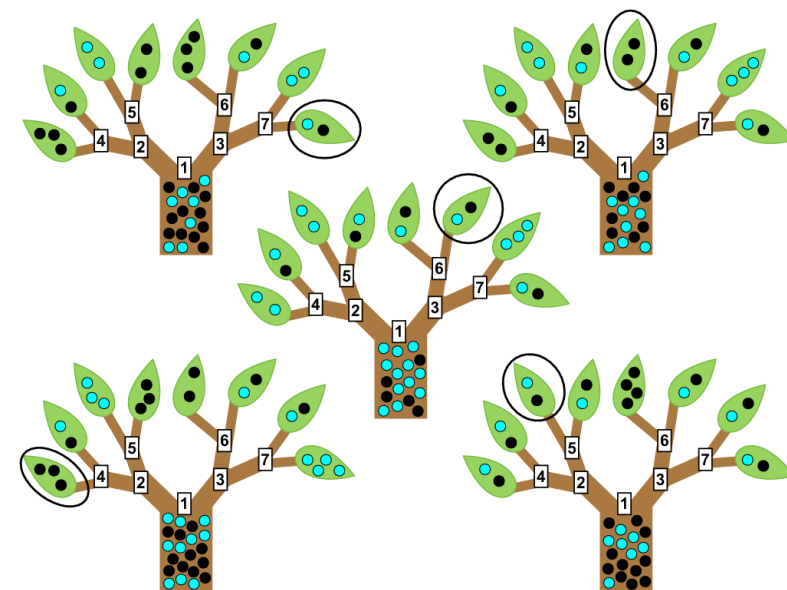
# Random Forests





# Random Forests

- Random forests are a slight trick to bagging that highly improves predictive power
- **Many trees do poorly because the stepwise greedy algorithm doesn't fully explore variable and parameter space**
- Random forests is like bagging, only each time a split in a tree is considered, a random selection of  $m$  predictors is chosen as split candidate
- A fresh set of  $m$  predictors is taken at each split.



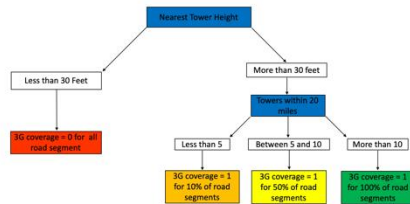
# Random Forest Model: Many Decision Trees

B bootstrap  
samples of data  
(~ 1000)



Bootstrap 1

Tree 1

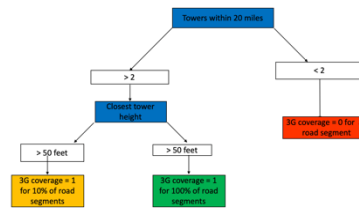


Vote if household is poor



Bootstrap b

Tree b



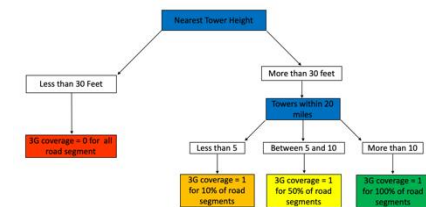
Vote if household is poor

...



Bootstrap 1000

Tree 1000



Vote if household is poor

# Estimating Random Forest Models Using “randomForest”

randomForest {randomForest}

R Documentat

## Classification and Regression with Random Forest

### Description

randomForest implements Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code) for classification and regression. It can also be used in unsupervised mode for assessing proximities among data points.

### Usage

```
## S3 method for class 'formula'
randomForest(formula, data=NULL, ..., subset, na.action=na.fail)
## Default S3 method:
randomForest(x, y=NULL, xtest=NULL, ytest=NULL, ntree=500,
             mtry=if (!is.null(y) && !is.factor(y))
               max(floor(ncol(x)/3), 1) else floor(sqrt(ncol(x))),
             replace=TRUE, classwt=NULL, cutoff, strata,
             sampsize = if (replace) nrow(x) else ceiling(.632*nrow(x)),
             nodesize = if (!is.null(y) && !is.factor(y)) 5 else 1,
             maxnodes = NULL,
             importance=FALSE, localImp=FALSE, nPerm=1,
             proximity, oob.prox=proximity,
             norm.votes=TRUE, do.trace=FALSE,
             keep.forest=!is.null(y) && is.null(xtest), corr.bias=FALSE,
             keep.inbag=FALSE, ...)
## S3 method for class 'randomForest'
print(x, ...)
```

- Key parameters in the randomForest package:
  - Mtry: number of variables to randomly select at each candidate node split
  - Ntree: number of regression or classification trees used to fit total random forest model

# Estimating Random Forest Model Using randomForest() package

```
library('randomForest')

rf_fit <- randomForest(poor_stat ~ .,
                      data = CR_train %>%
                        select(-household_ID) %>%
                        mutate(poor_stat =
                              as.factor(poor_stat)),
                      type = classification,
                      mtry = 3,
                      na.action = na.roughfix,
                      ntree = 100,
                      importance = TRUE)
```

- Must specify formula and dataset to use per usual
- Additionally must specify “mtry” the number of variables to sample (randomly) for each node
- Missing values will cause an error and this rough fix replaces them with median values
- Ntree specifies the number of trees in the random forest

# Random Forest Model Object

```
> print(rf_fit)

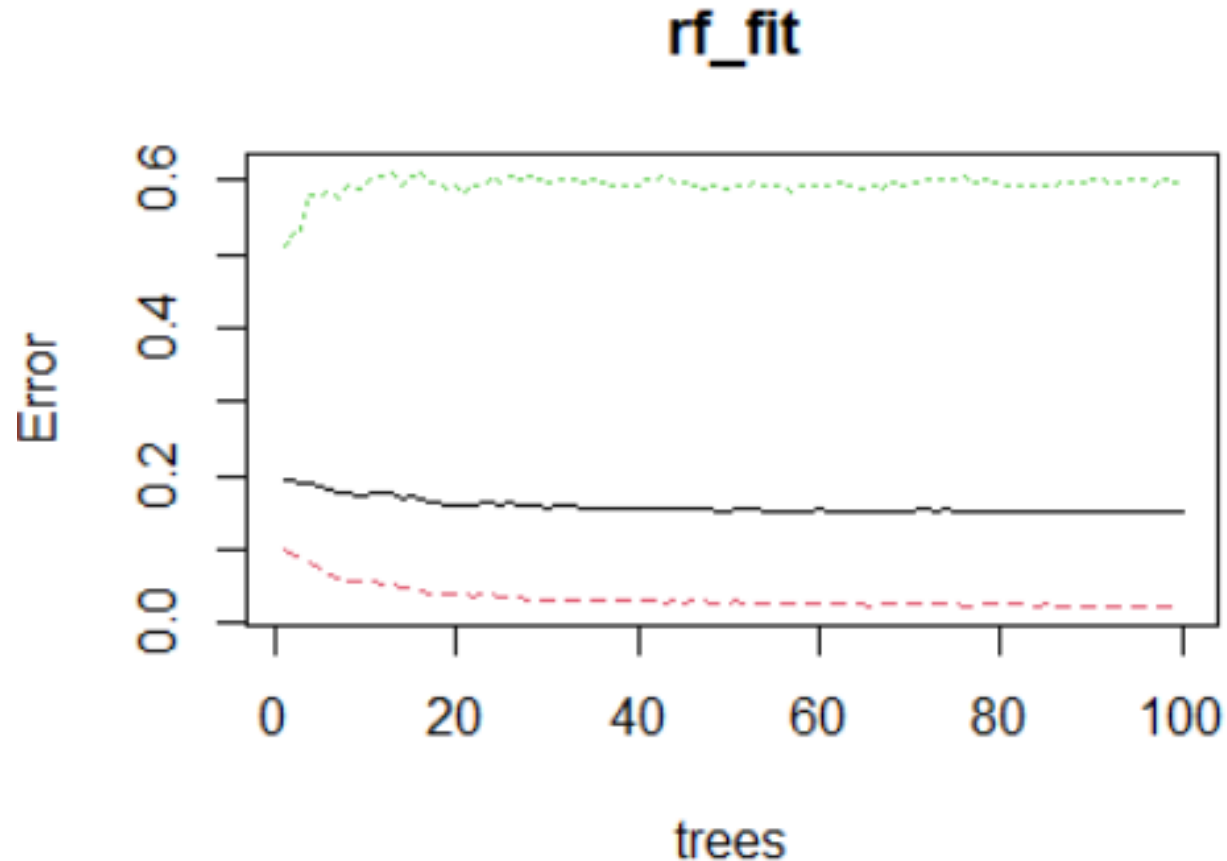
Call:
randomForest(formula = poor_stat ~ ., data = CR_train %>% select(-household_ID) %>% mutate(poor_stat = as.factor(poor_stat)), type = classification, mtry = 3, ntree = 100, importance = TRUE, na.action = na.roughfix)

Type of random forest: classification
Number of trees: 100
No. of variables tried at each split: 3

OOB estimate of error rate: 14.93%
Confusion matrix:
      0   1 class.error
0 3390  76 0.02192729
1  589 400 0.59555106
```

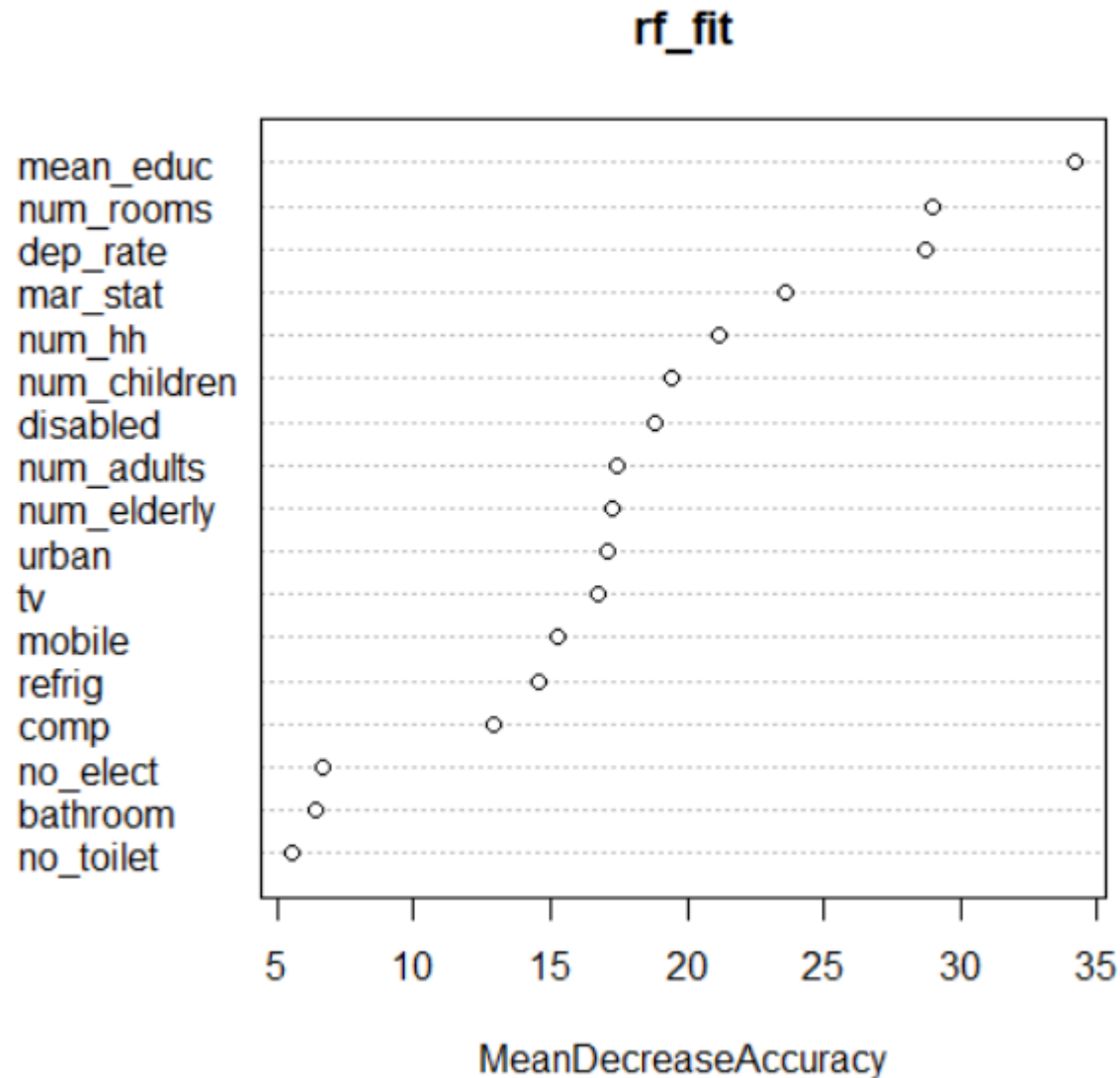


plot() function against rf object – 100 trees, mtry = 3



- Green is error rate for positive class, red is error rate for negative class (not poor), and black is overall error rate (all out of bag error)
- Error seems to stabilize at 50-100 trees, so we only need around 100 trees for this prediction problem

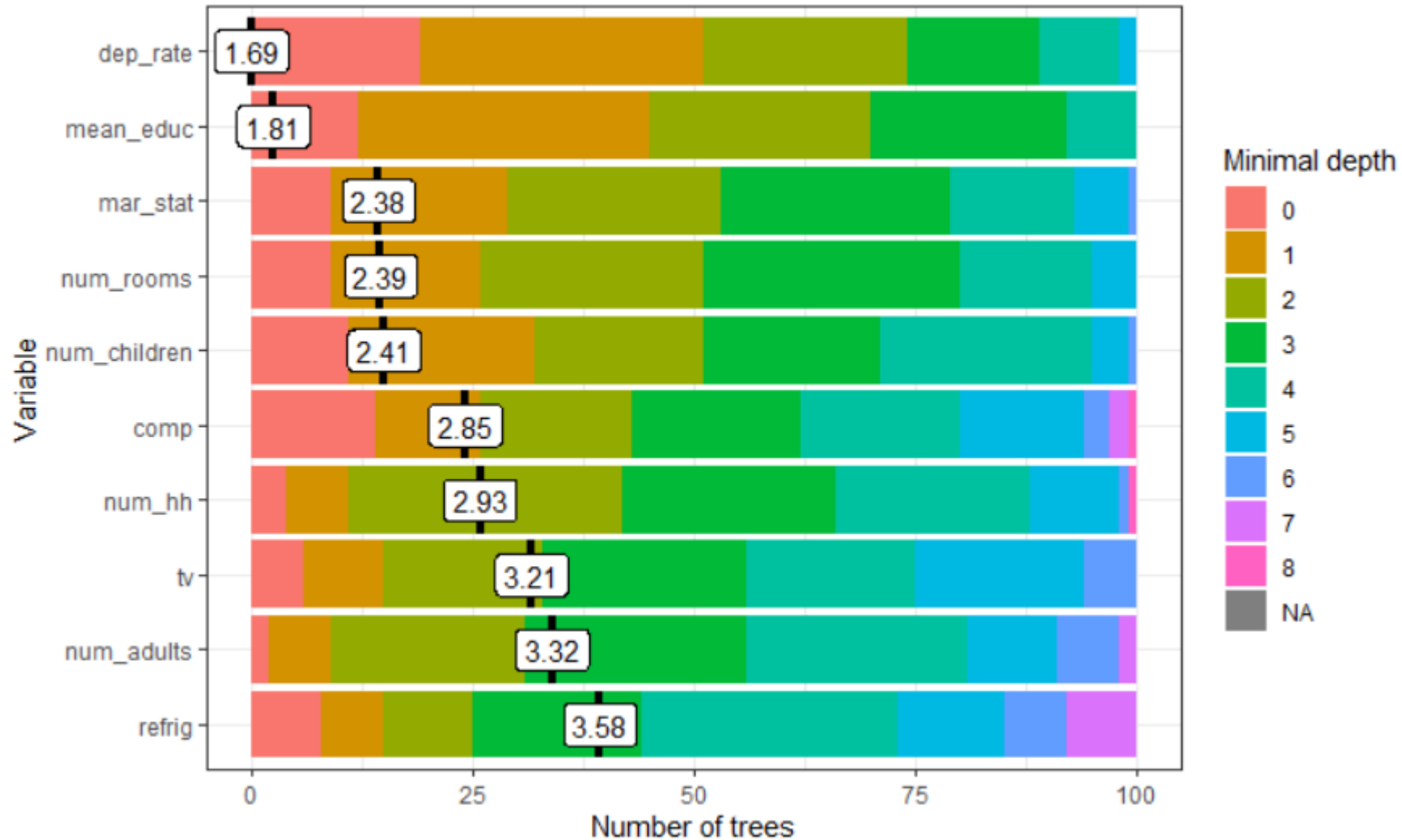
# Which Variables Matter Most? Variable Importance



- Which variables are most important for the random forest model?
- One way of assessing importance is using **variable permutation**. This replaces a given variable (sequentially) with random noise (e.g. `rnorm(., 0,1)`) and re-estimates the model.
- Logic: more “important” variables result in worse models when these variables are absent (or are random noise)
- `varImpPlot(rf_fit)` plots these importance measures
- Mean educ is most important, followed by number of rooms, then dep rate

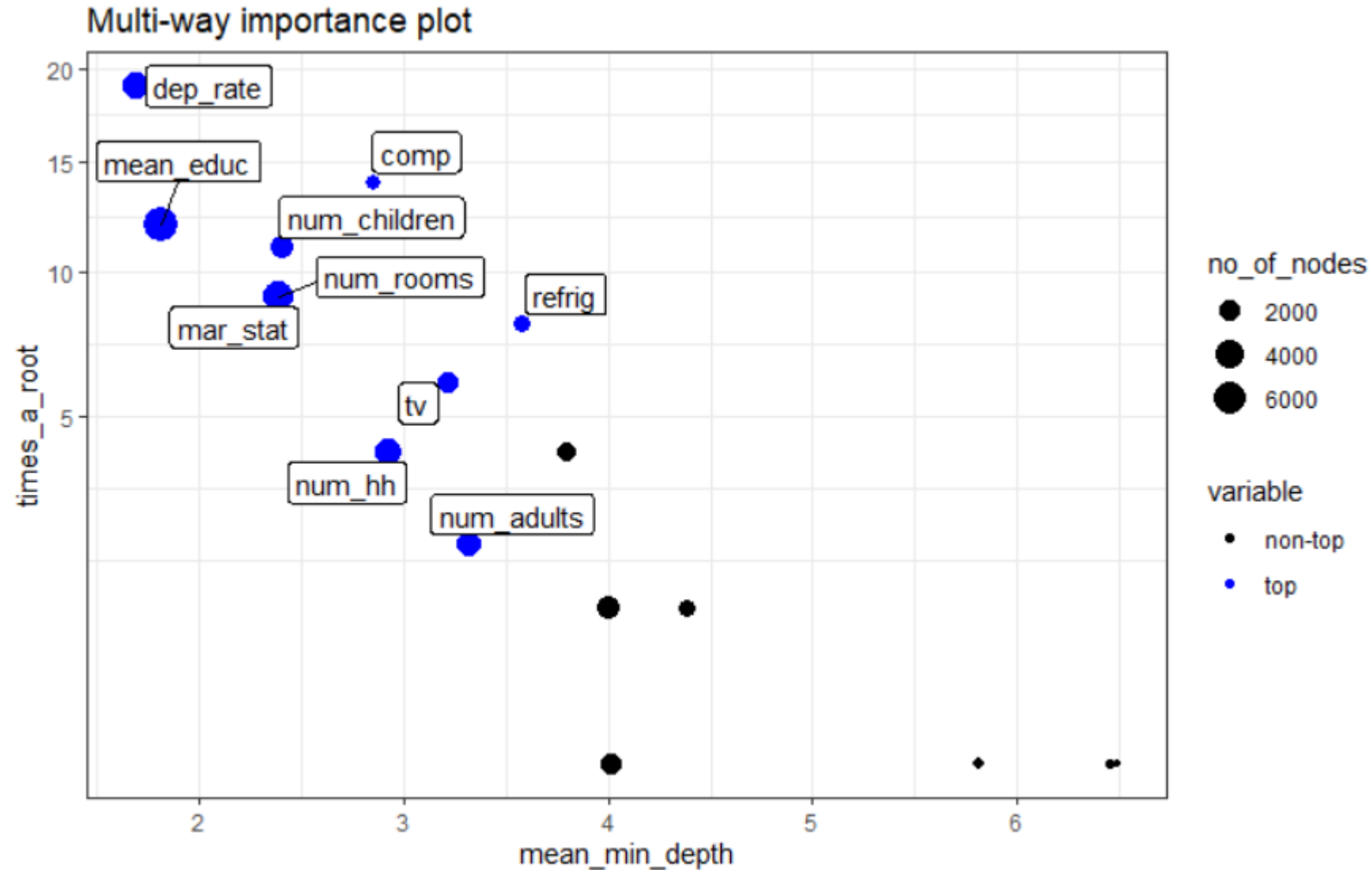
# randomForestExplainer

Distribution of minimal depth and its mean

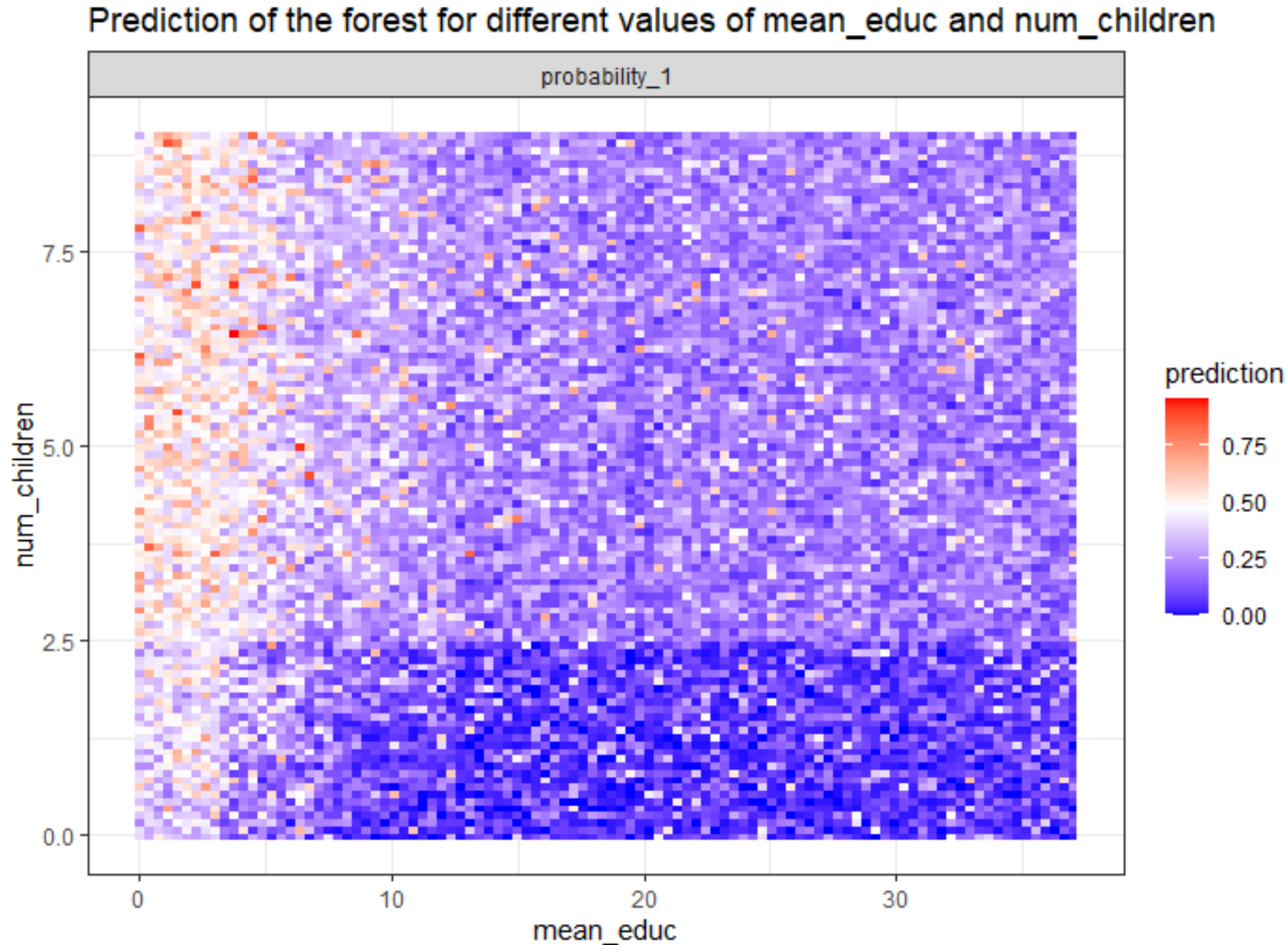


- Random forest explainer is a package that has many useful functions to explain and open up the random forest black box.
- This plots the average depth of trees in the ensemble that use these variables

# randomForestExplainer



# randomForestExplainer



- We can see how the random forest classifies households with different characteristics of average education and number of children

# Cross-Validating to Select mtry in caret

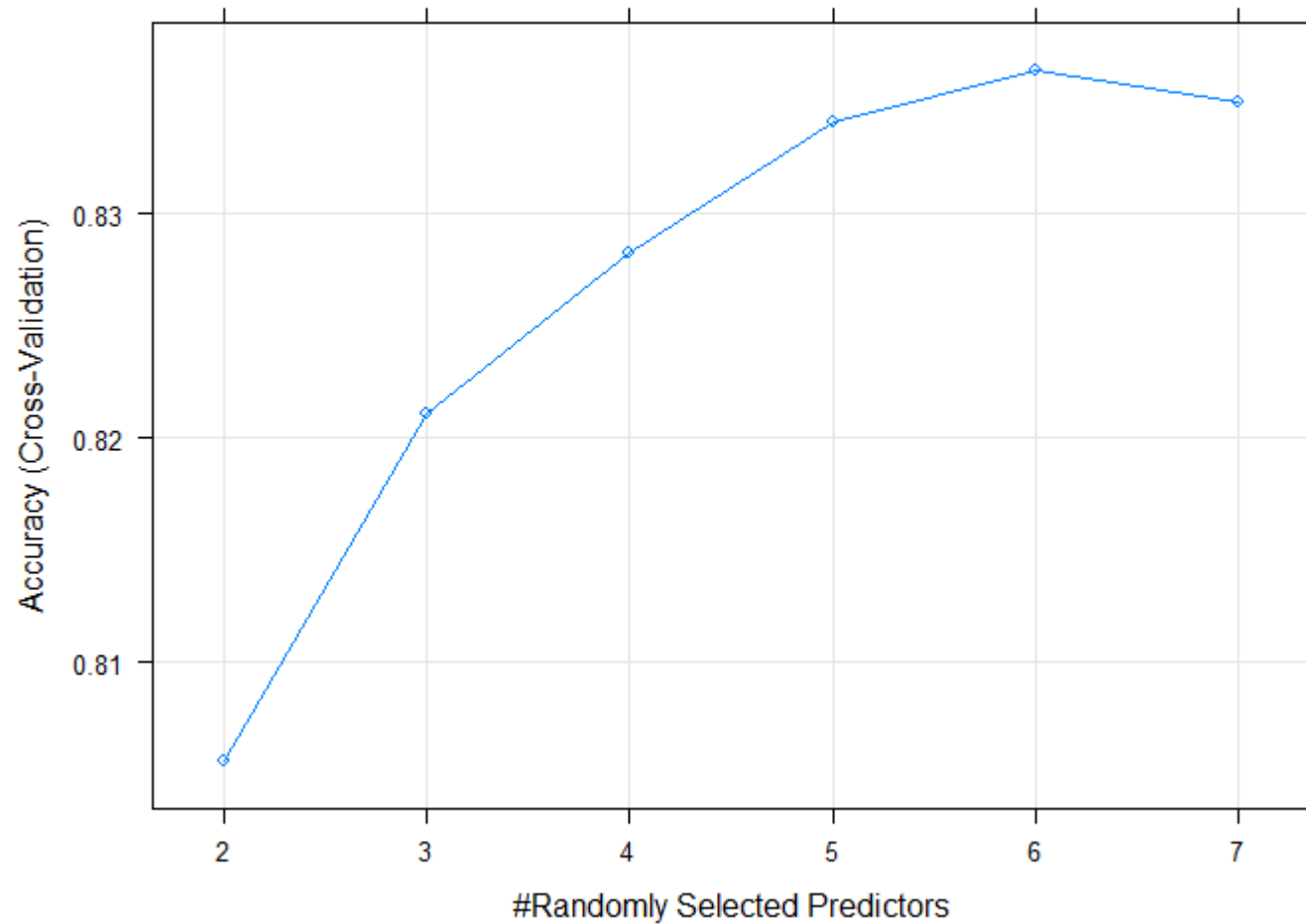
```
library('caret')

rf_caret <-
  train(poor_stat ~ urban + num_children + comp + no_toilet
        + dep_rate + mobile + mean_educ,
        CR_train %>%
          select(-household_ID) %>%
          mutate(poor_stat =
                 as.factor(poor_stat)),
        method = "rf",
        metric = "Accuracy",
        tuneLength = 10,
        trControl = trainControl(method = "cv",
                                   number = 5,
                                   verbose = TRUE))
```

- caret is a great machine learning package that holds many models.
- It handles many common machine learning tasks such as cross-validation to select optimal parameters



# Cross-Validating to Select mtry in caret



- It seems we should set `mtry = 6`!