An Introduction to Trees and Random Forests

James Owers, University of Edinburgh

1st February 2015

Leo Breiman (1928-2005)

- And why should I bother listening…

Why Random Forests

Flexible and simple to implement

- Takes any data type as-is: no data transformations required
- R and Python off-the-shelf implementations are good
- Can be used to fill in missing values
- Model validation is inbuilt

Easy to understand and interpret

- Variable importance is a by product of the model
- Trees reveal the model's thought process

//

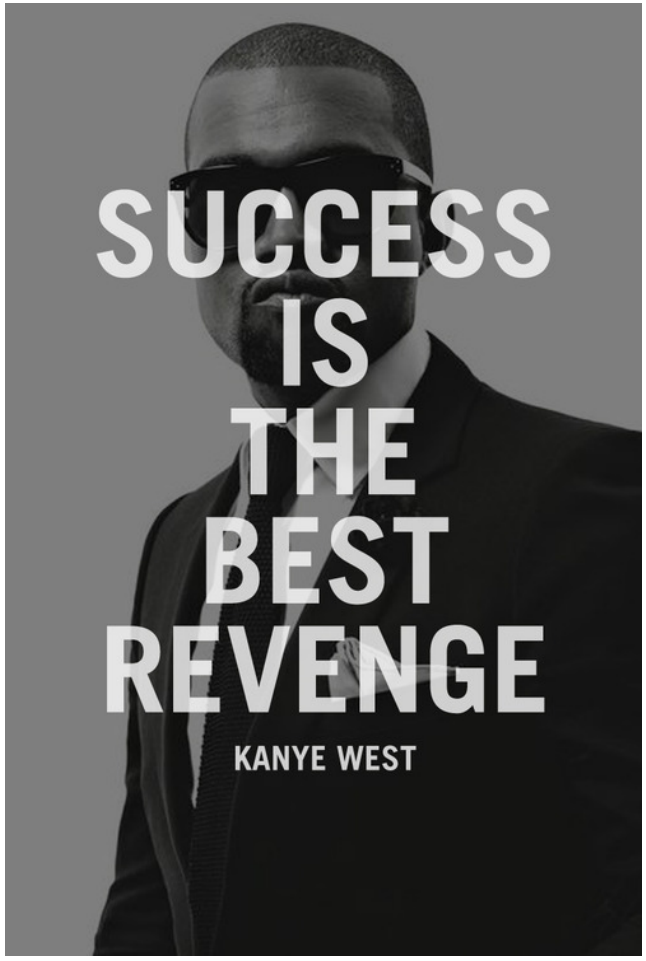
Fast & Accurate

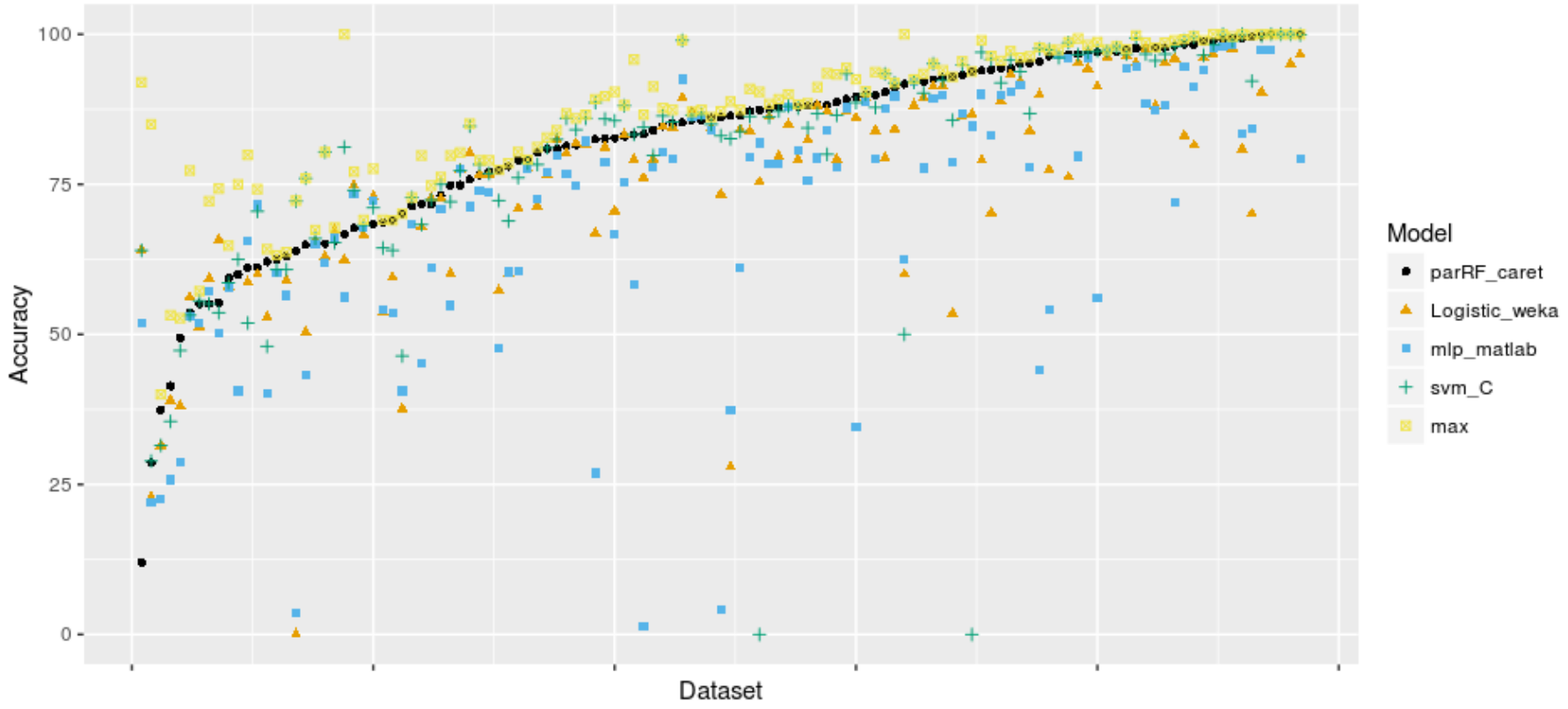
- Easily parallelisable
- Wins Kaggle competitions
- Used by big companies e.g. Microsoft in the Kinect
- Active research c.f. Decision Forrests (Criminisi and Shotton)
- They are the single greatest living model

Jokes aside, when 179 classifiers were trained on all 121 datasets in the UCI Machine Learning repository, Random Forests swept the floor:

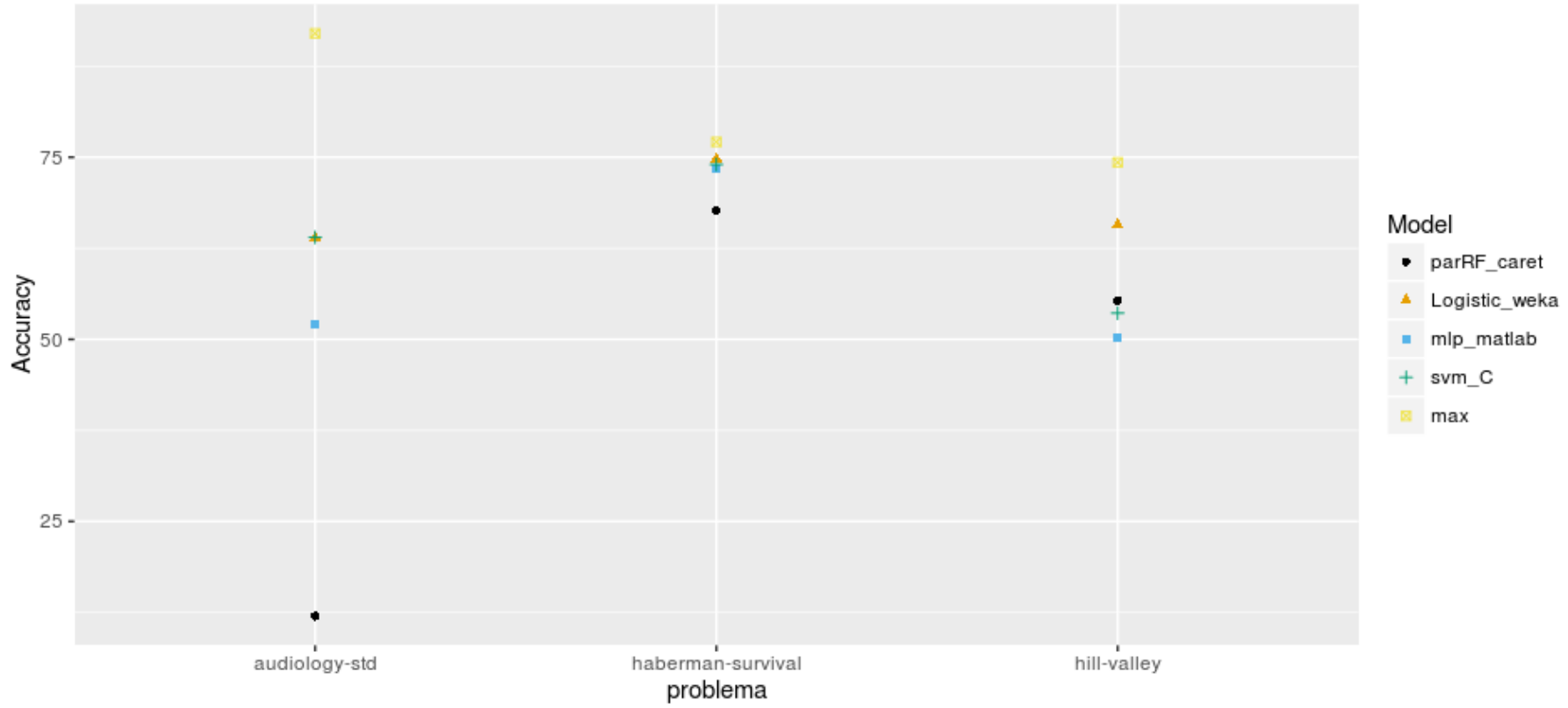
- the **highest average accuracy**
- in the top 10% of classifiers for 84% of the datasets

(Source: Fernandez-Delgado et al 2014 - Journal of Machine Learning Research)

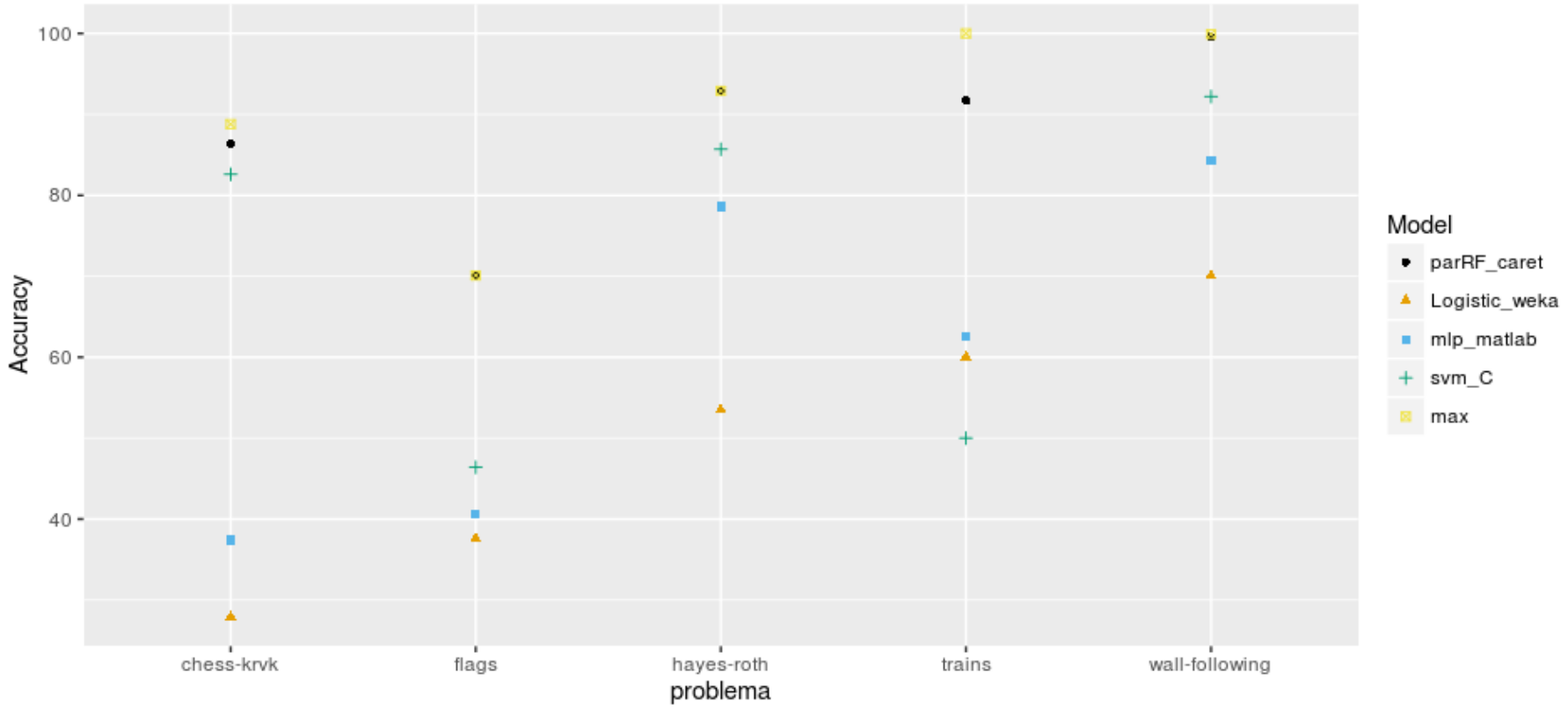

SUCCESS
IS
THE
BEST
REVENGE
KANYE WEST



The 121 datasets ordered by Random Forest accuracy and compared with Logistic Regression (multiclass), MLP (one hidden layer), and SVM (gaussian kernel)



Selected datasets where Random Forest performs poorly against Logistic Regression (any where it is at least 5 percentage points worse)



Selected datasets where Random Forest performs well against Logistic Regression (any where it is at least 25 percentage points better)

Selected datasets where RF performs poorly

Dataset	RF_vs_LR	N	features	classes	pct_maj_class	Desc
audiology-std	-52	226	59	18	26	Diagnose hearing ailment
haberman-survival	-7.1	306	3	2	74	Predict if the patient will survive longer than 5 years
hill-valley	-10	606	100	2	51	Determine whether you are on a hill or in a valley depending on x1-x99 and y information

Selected datasets where RF performs well

Dataset	RF_vs_LR	N	features	classes	pct_maj_class	Desc
chess-krvk	58	28056	6	18	16	Predict nr moves till win (interestingly chess-krvkp solved well with logistic)
flags	32	194	28	8	31	Predict country religion from flag and country stats
hayes-roth	39	132	3	3	39	Predict class from entirely categorical data
trains	32	10	32	2	50	Determine concise decision rules distinguishing trains traveling east from those traveling west
wall-following	30	5456	24	4	40	Classify which way a robot should move to avoid collisions (experiment to show nonlinear)

Decision Trees

What are they

- A succession of binary decisions
- Predictions are made at each leaf
- Basically the design of an annoying facebook quiz
- *"Which Lord of the Rings character are you?!"*

Why use them

- Out-of-the-box magic
  - Takes a mix of categorical and continuous unscaled data
- Interpretable results
- Quick to train

Learning the splits

Learning method for the CART algorithm (binary trees)

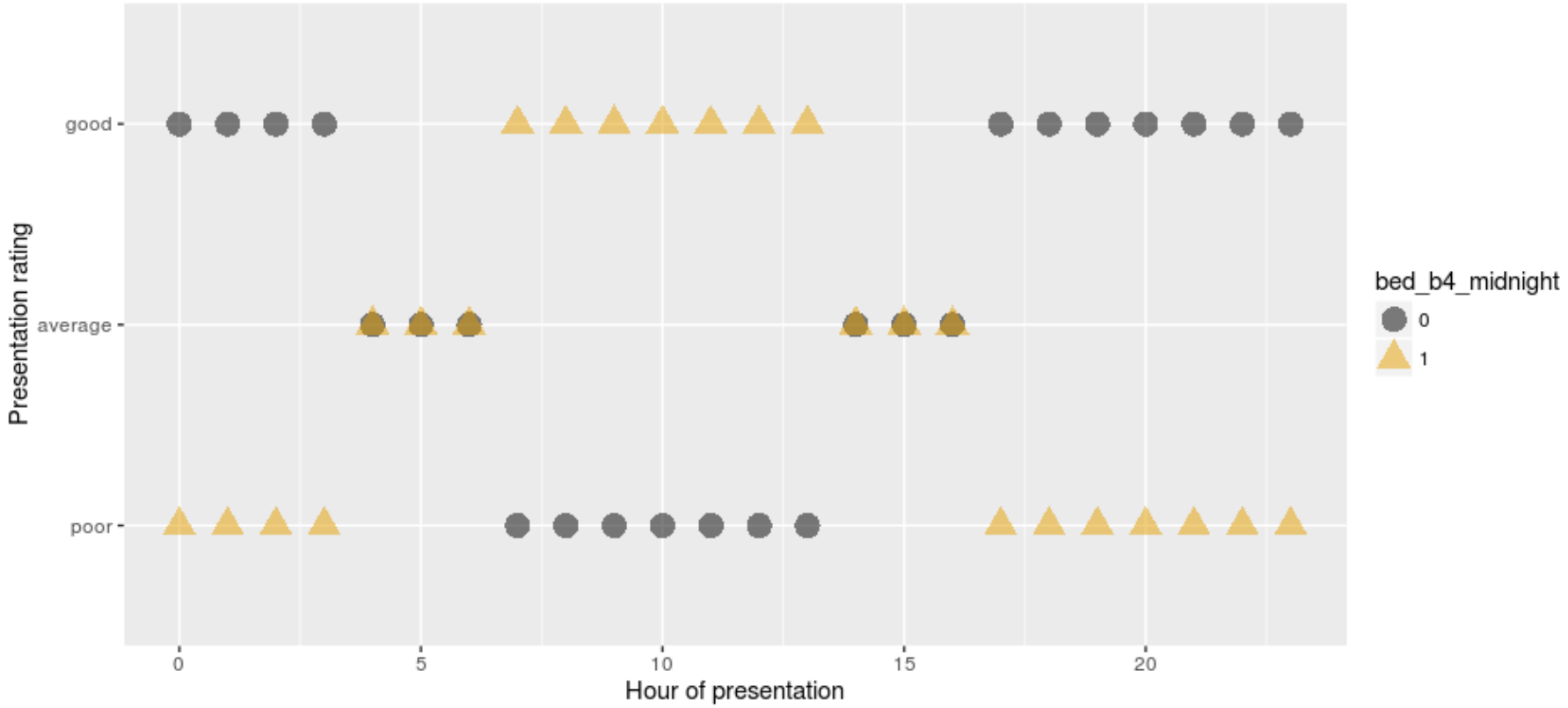
- Categorical features with C categories: $2^{C-1} - 1$ possible splits
- Continuous or ordinals features with K values: $K - 1$ possible splits

Algorithm:

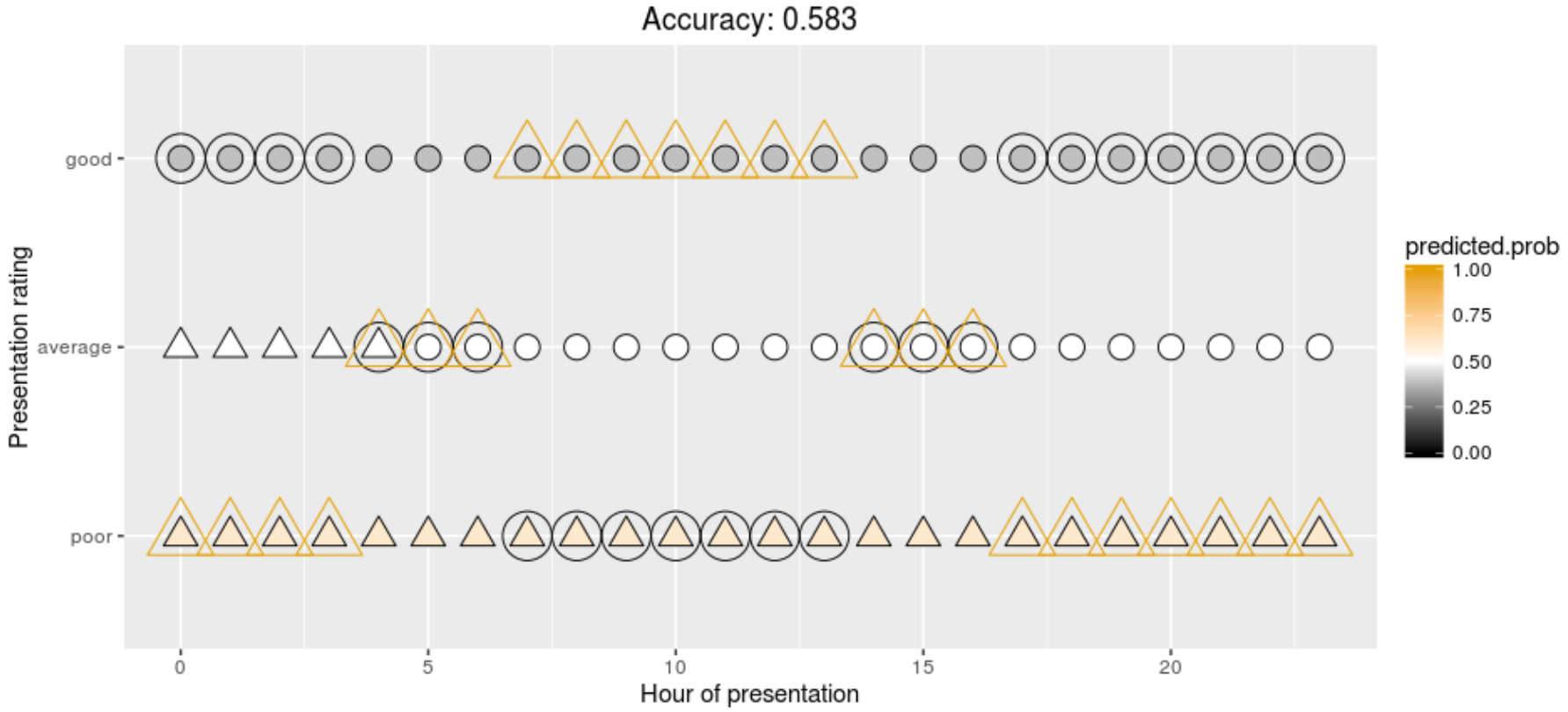
- Evaluate the "splitting criterion" at all possible splits
- Pick the best one
- Repeat until "splitting criterion" gets worse, or child is pure

Consequence: will never think outside the box

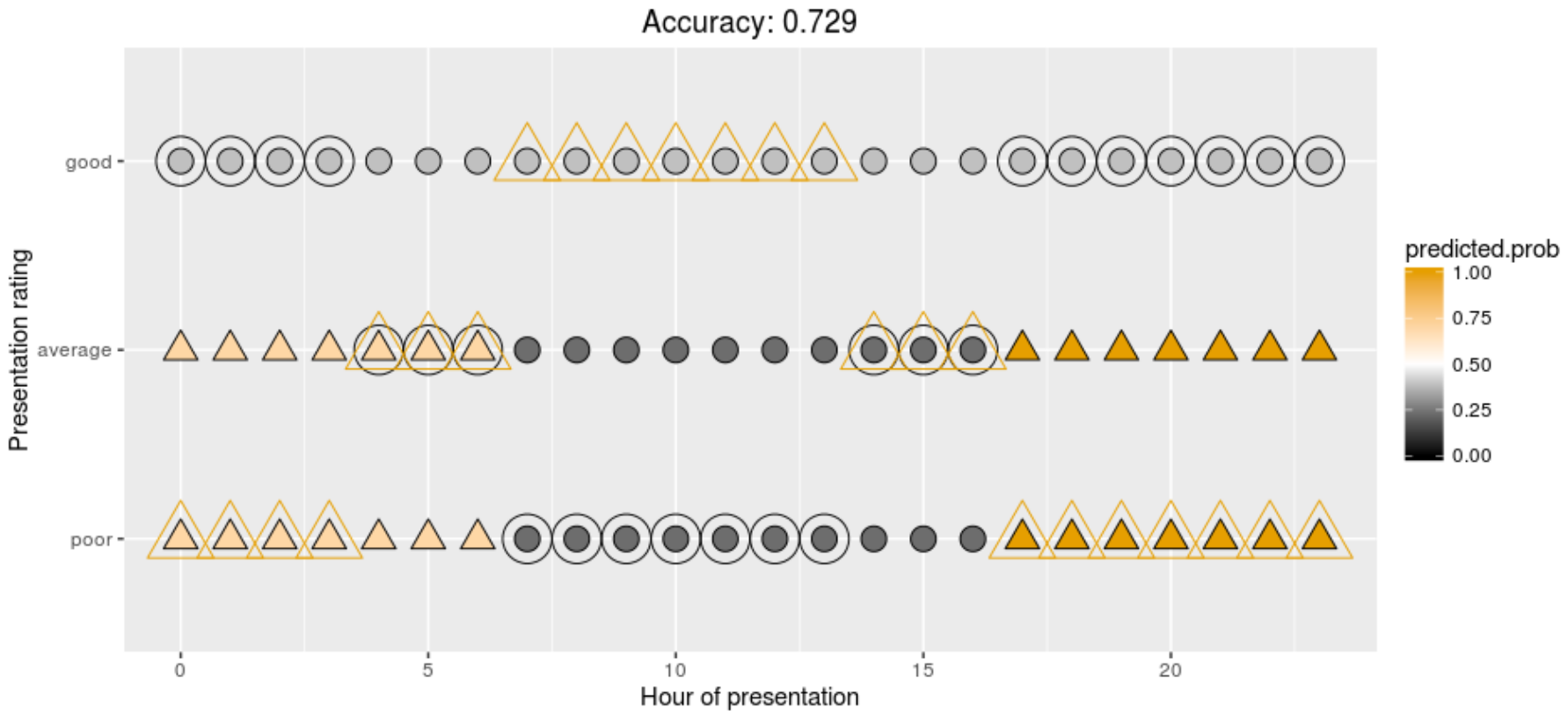
Toy Example - Predicting Bedtime



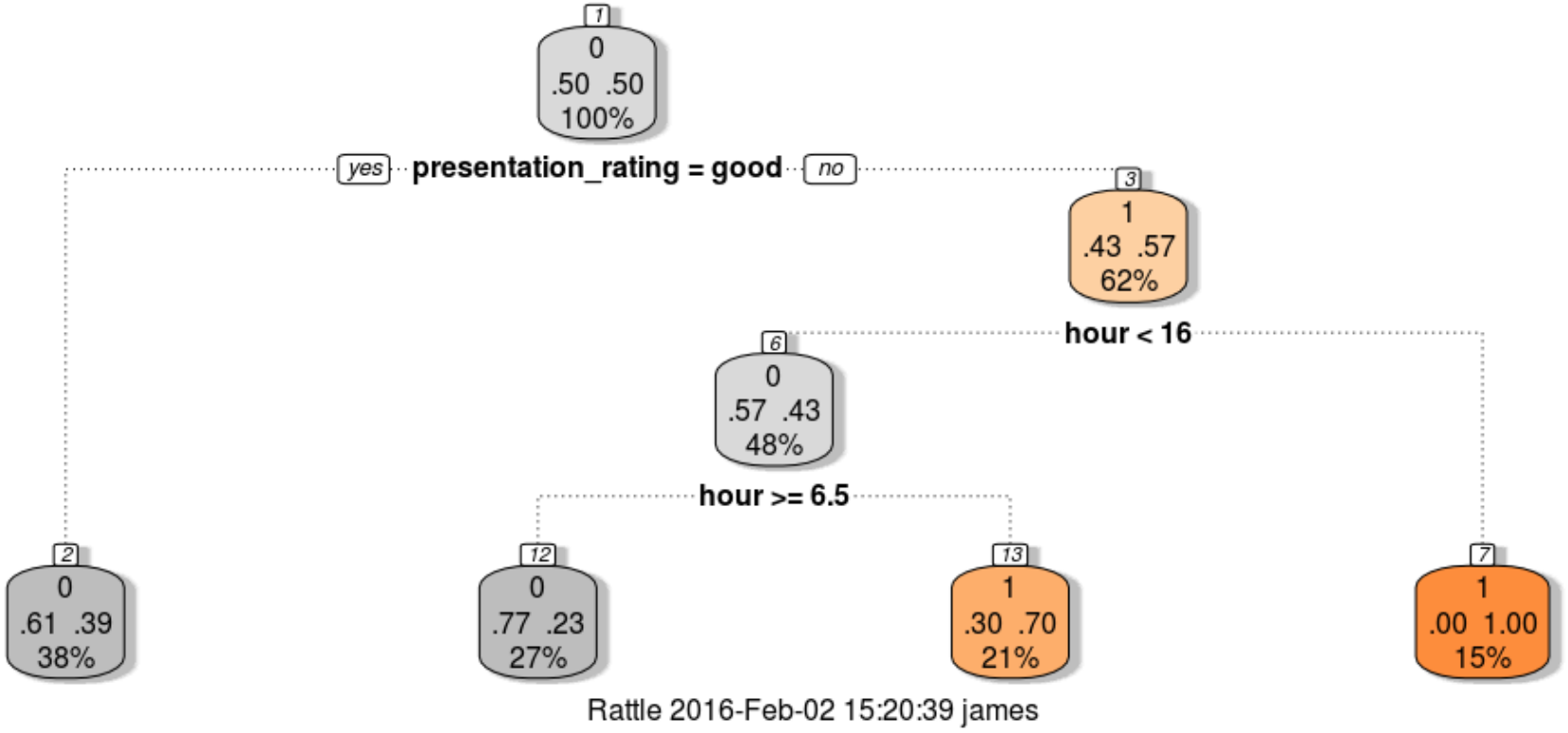
Logistic Regression Fit



Decision Tree Fit



The learned decision tree



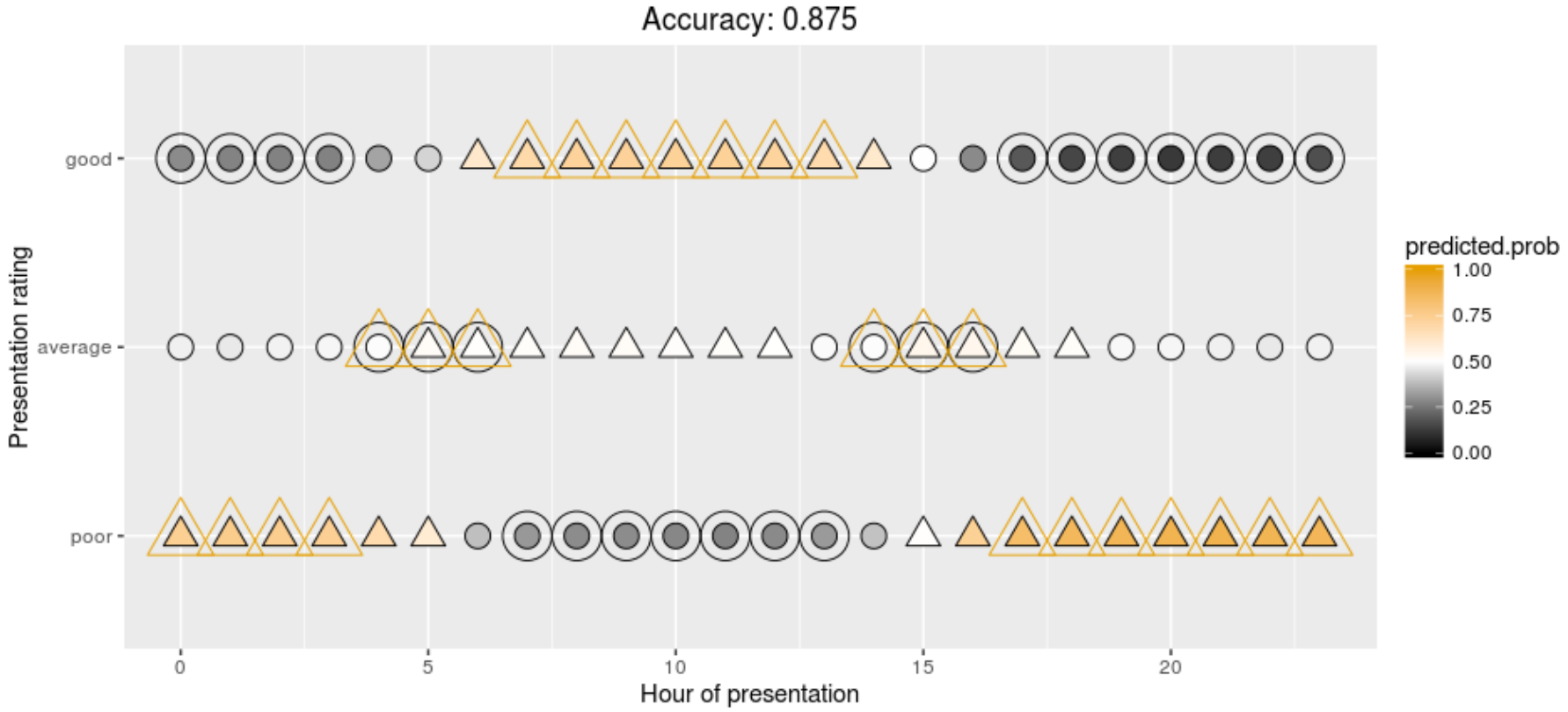
Main advantage of Decision Trees

- Can model interactions
- No need to generate more features in your data

Main issues with Decision Trees

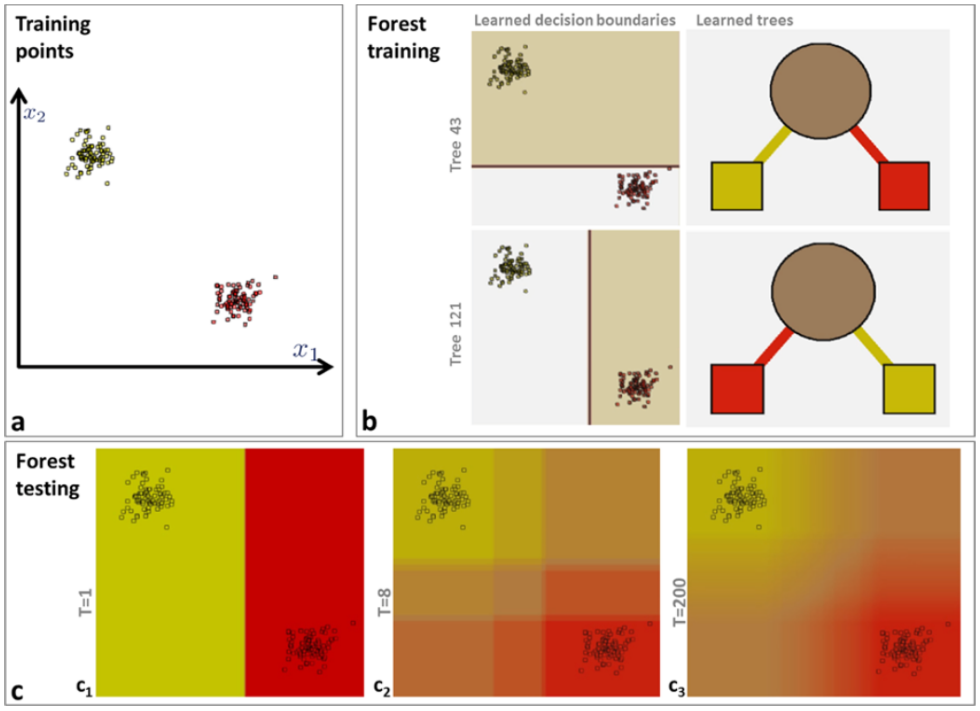
- Baised
- Splits space into boxes

Random Forrest Fit



Random Forests

Forests - Averaging many trees



From Criminisi & Shotton 2013 Decision Forests for Computer Vision and Medical Image Analysis

Reducing the variance

- Bias variance tradeoff in action

Randomise splitting

- Choose a subset of features to consider
- Choose the feature in the subset with the best splitting critereon

Randomise training data

- Draw a different bootstrap sample for each tree
- Each tree is more biased than the original
- Ensemble of weak learners is better than 1 strong

Tuning parameters

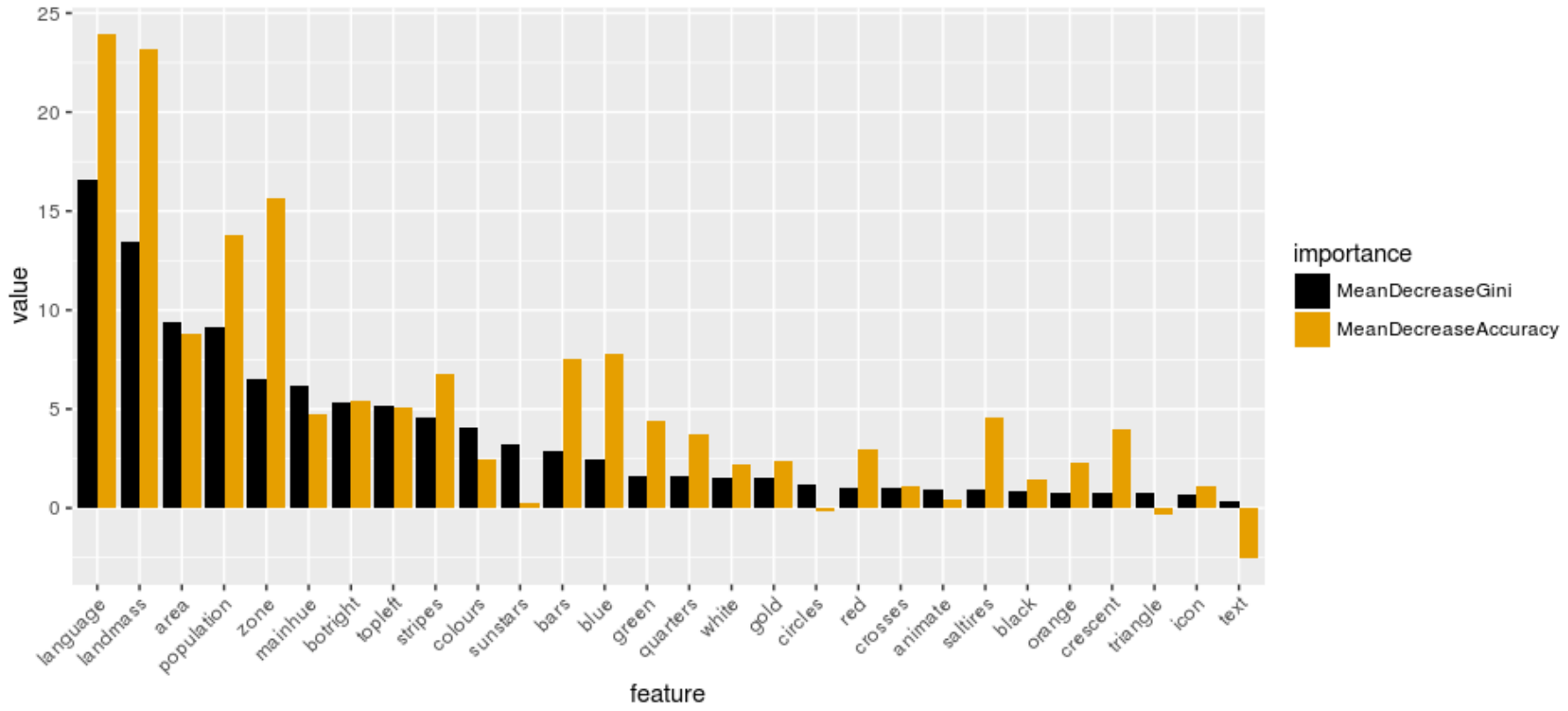
- Complexity of trees (depth etc.)
- Number of features considered at each split
- ~~Random Forrests dont overfit~~ number of trees not a tuning parameter

Feature importance

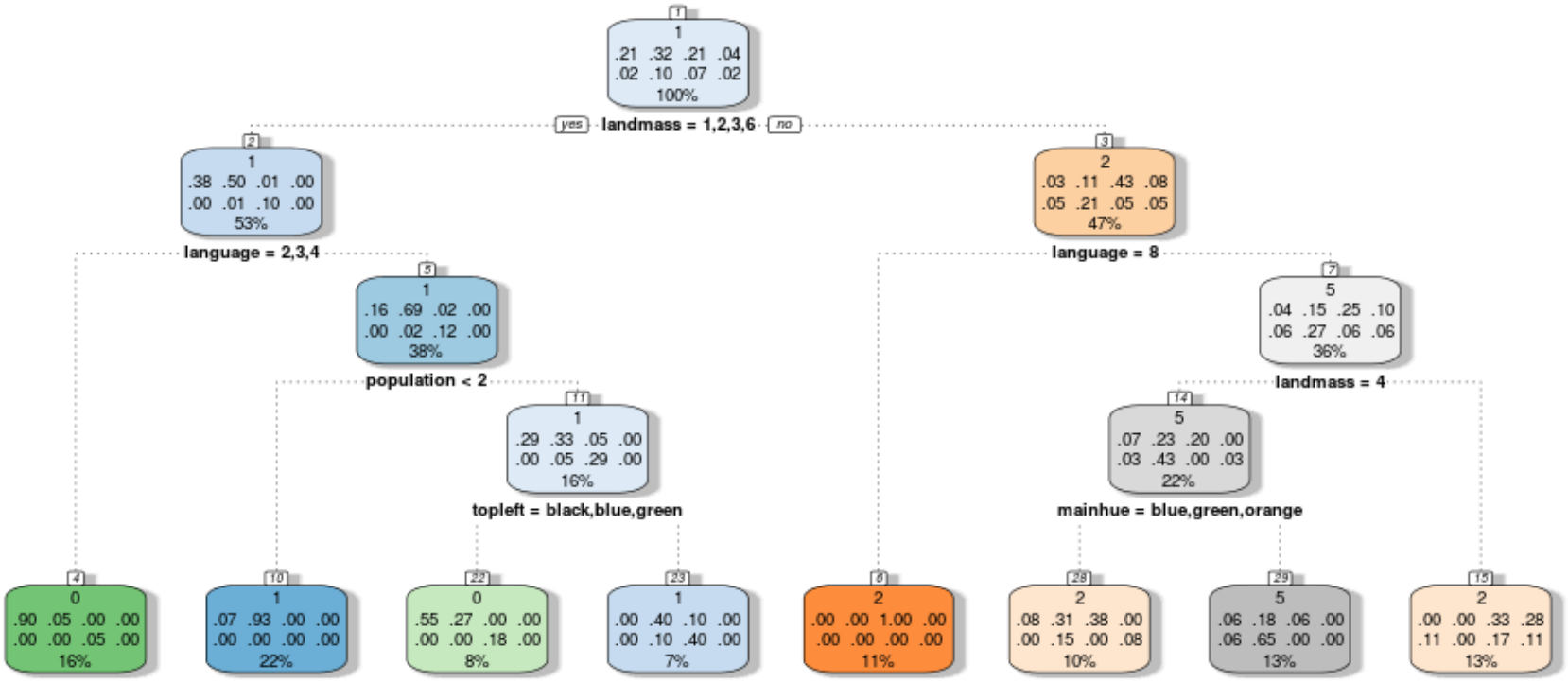
- Average out the improvement in the splitting critereon
- Can even work per output class
- Doing this for regression involves permutation

Example: Flags data

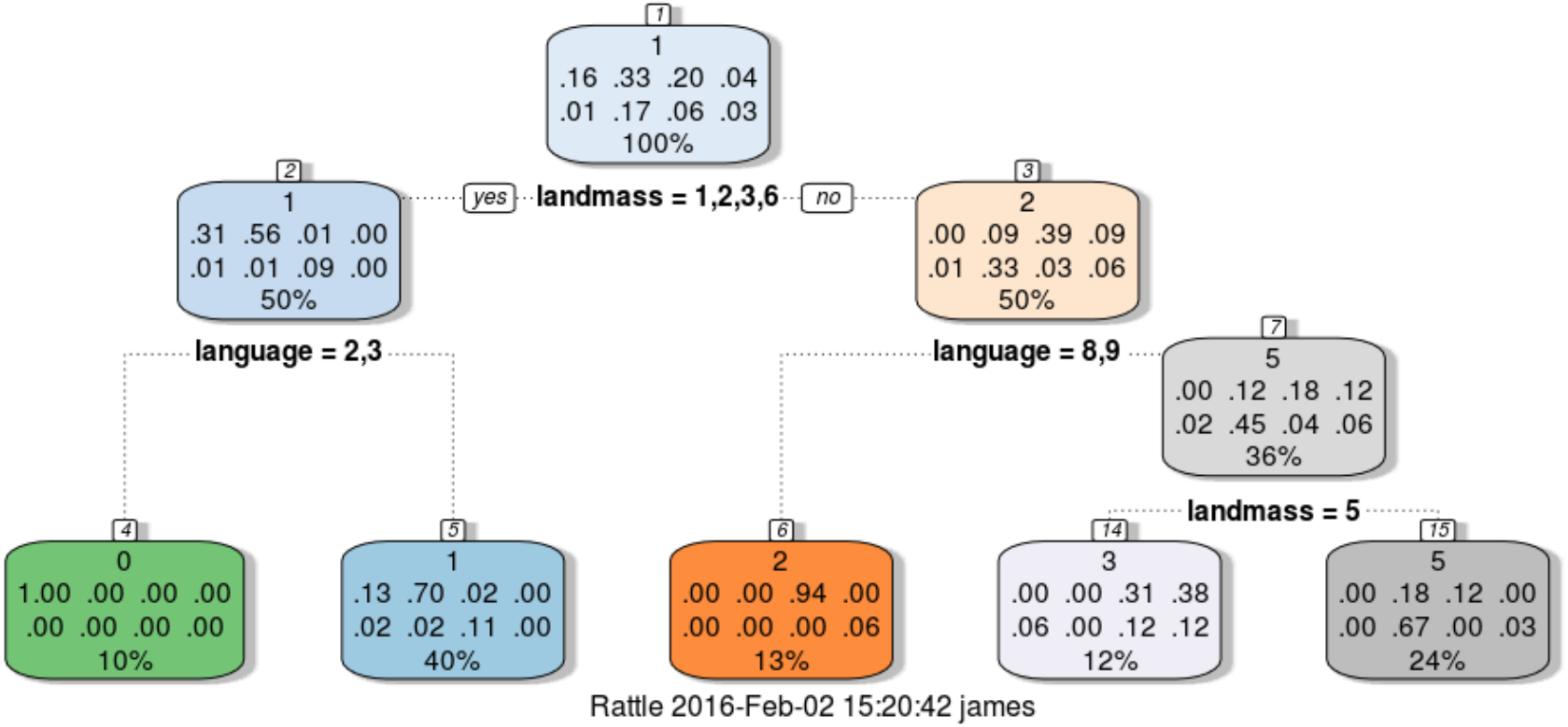
- 28 features: country statistics and flag information
- Outcome: religion - 8 classes

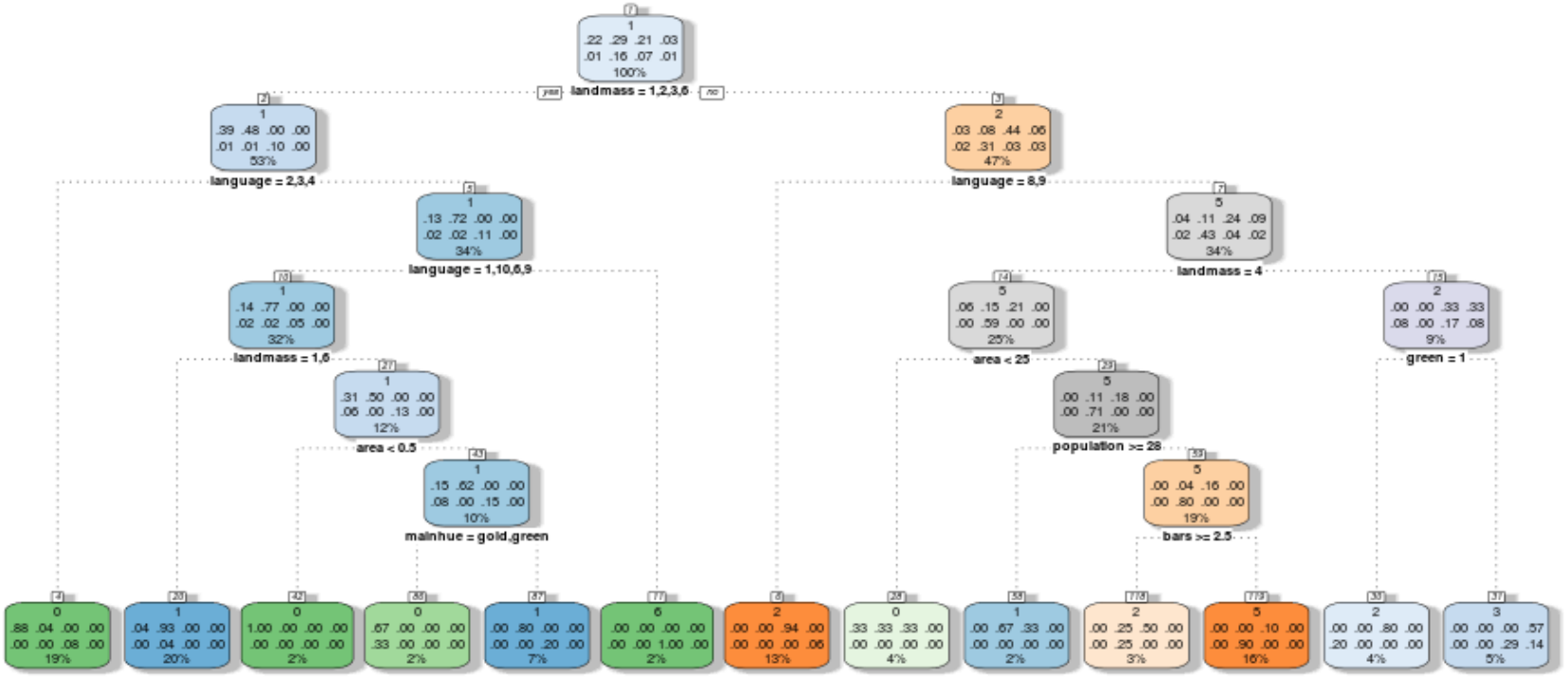


Feature importance for an out-of-the-box Random Forest trained on `Flags`



Rattle 2016-Feb-02 15:20:42 james


1
1
.16 .33 .20 .04
.01 .17 .06 .03
100%
yes
landmass = 1,2,3,6
no
2
1
.31 .56 .01 .00
.01 .01 .09 .00
50%
3
2
.00 .09 .39 .09
.01 .33 .03 .06
50%
language = 2,3
language = 8,9
7
5
.00 .12 .18 .12
.02 .45 .04 .06
36%
landmass = 5
4
0
1.00 .00 .00 .00
.00 .00 .00 .00
10%
5
1
.13 .70 .02 .00
.02 .02 .11 .00
40%
6
2
.00 .00 .94 .00
.00 .00 .00 .06
13%
14
3
.00 .00 .31 .38
.06 .00 .12 .12
12%
15
5
.00 .18 .12 .00
.00 .67 .00 .03
24%
Rattle 2016-Feb-02 15:20:42 james



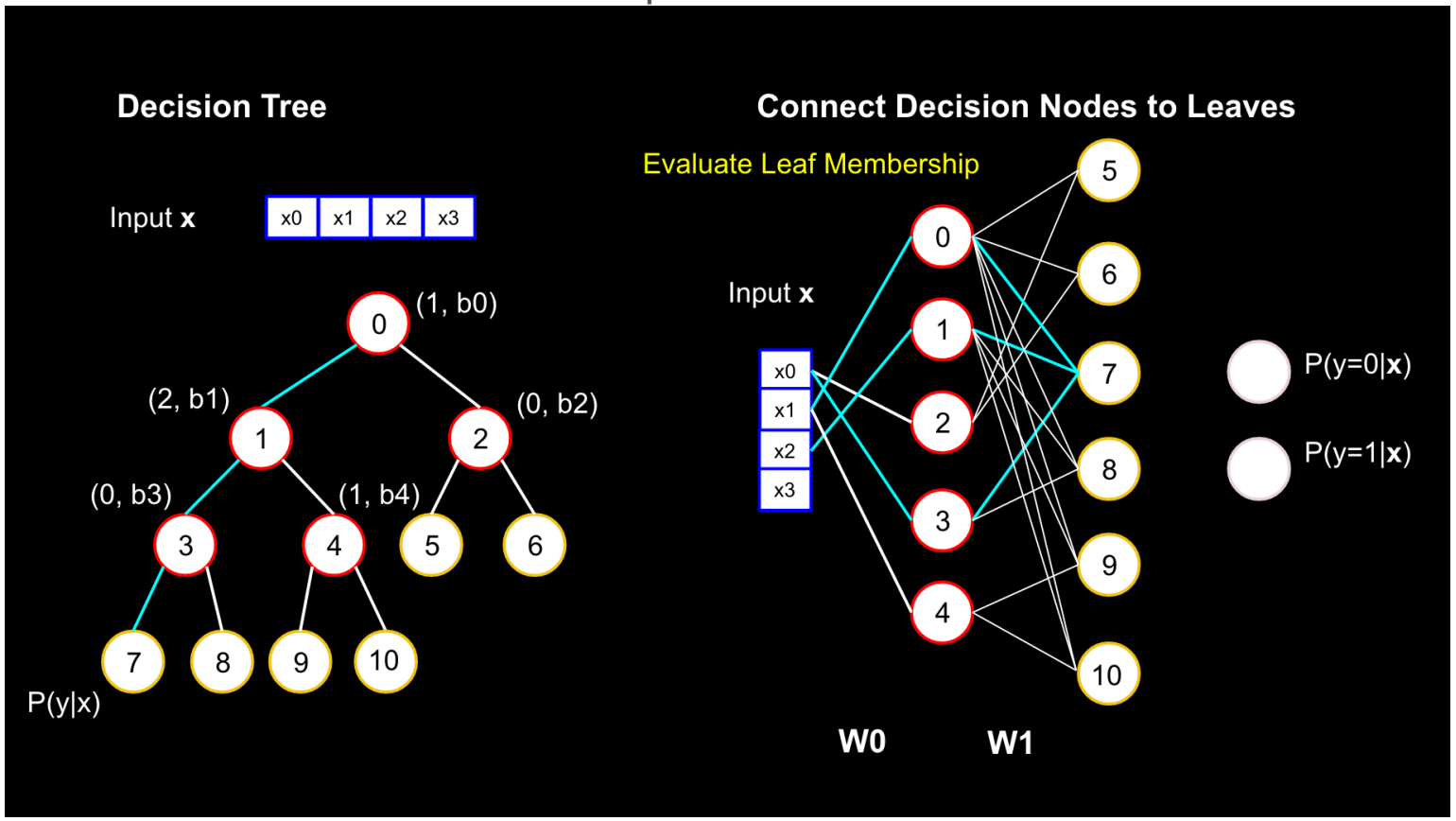
Rattle 2016-Feb-02 15:20:42 james

Tricks

- Non axis aligned splits e.g. kernalised decision forests
- Dealing with class imbalance with stratified sampling
- Missing value imputation by recursion initialising at the mean value
- Initialisation for a neural network (Welbl 2014) (Kontschieder 2015 @ ICCV)
- Feature factories - parametrise features

RF to NN

Oisin Mac Aodha - slide from presentation in 2015



Thank you for listening