Can’t see the trees for the forest…

28th June 2019

Emma Mills

Up – Sell: Many predictors?

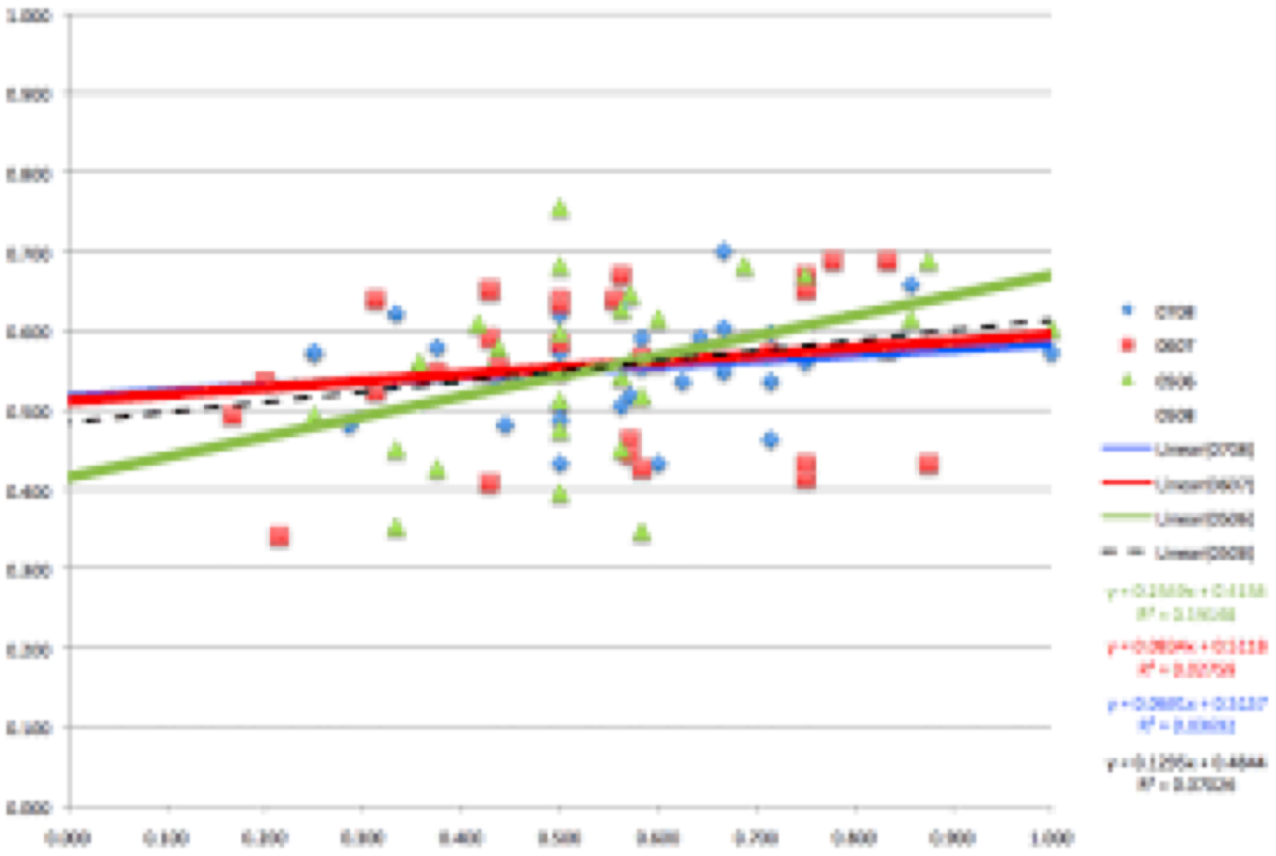


Image taken from analyticstraining.com

Up - Sell

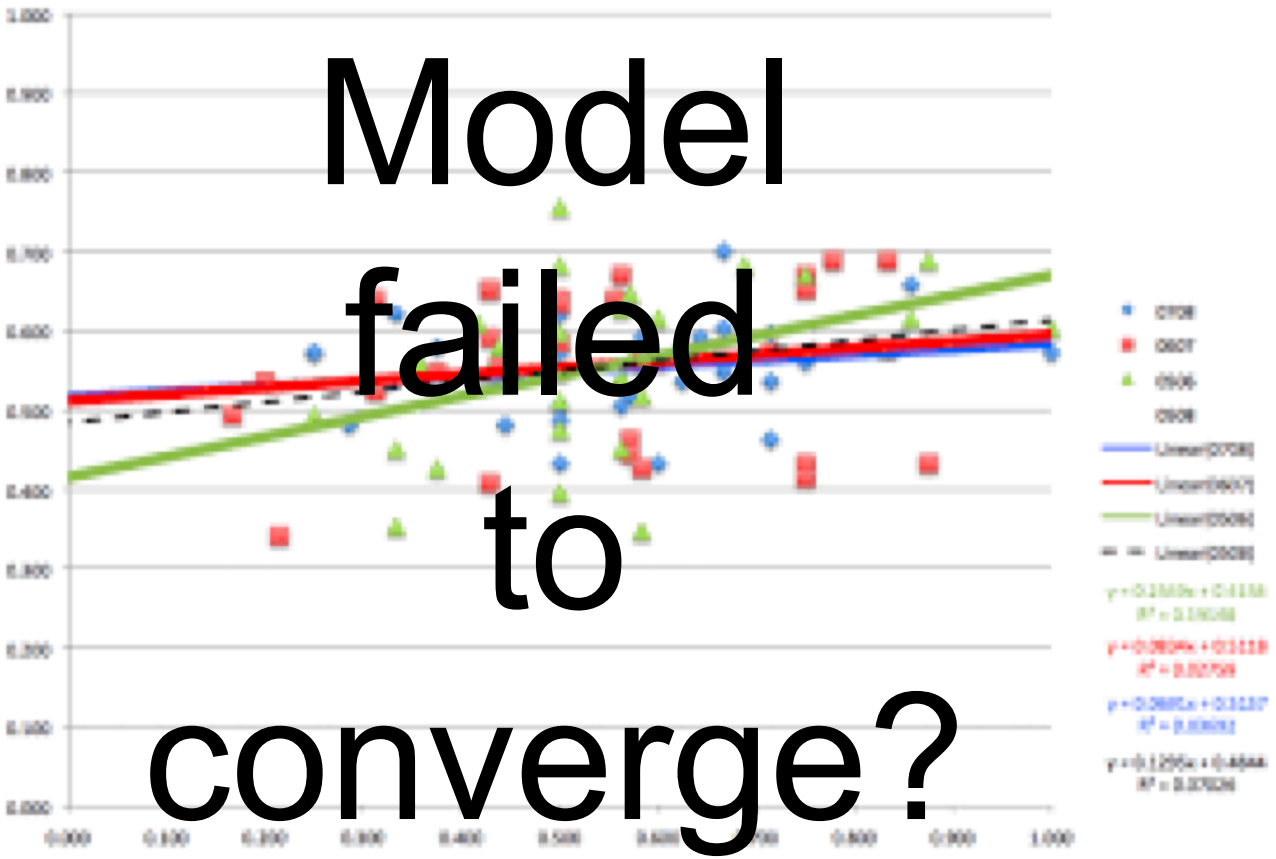


Image taken from analyticstraining.com

Up – Sell: Multi-collinearity?

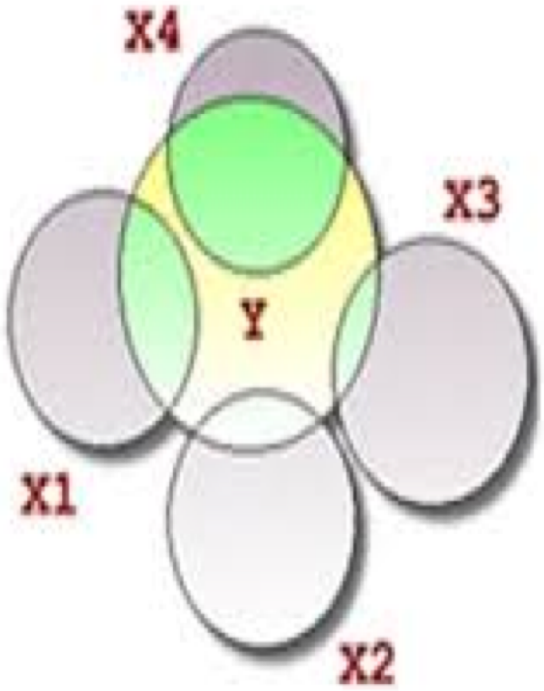


Image taken from analyticstraining.com

Up - Sell

Model almost singular?

Image taken from analyticstraining.com

Up-sell: Function of the outcome?

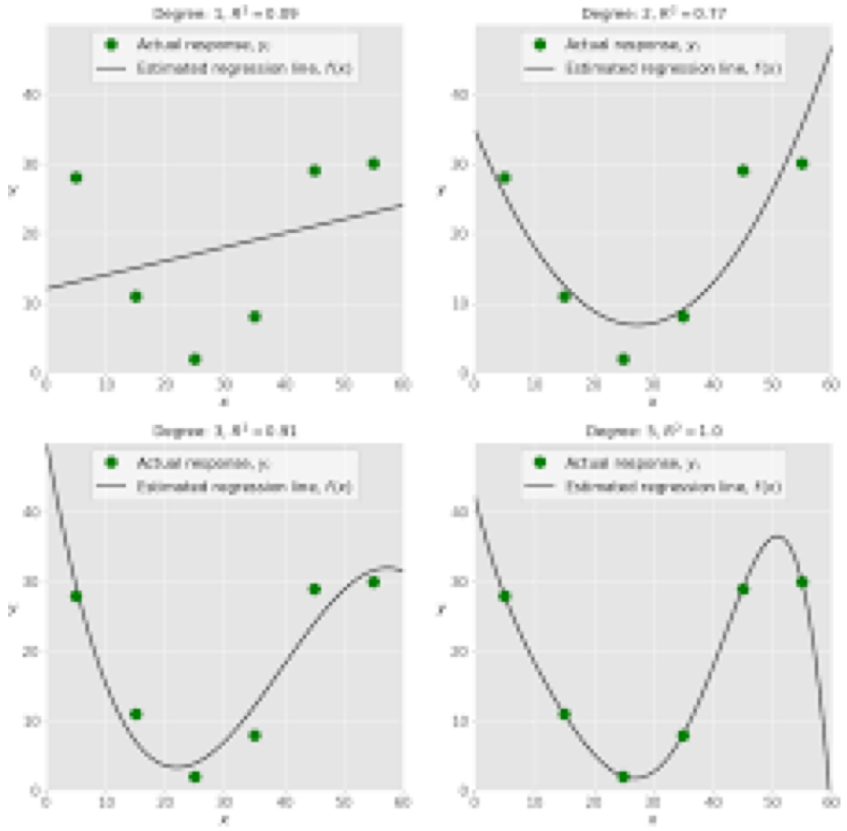
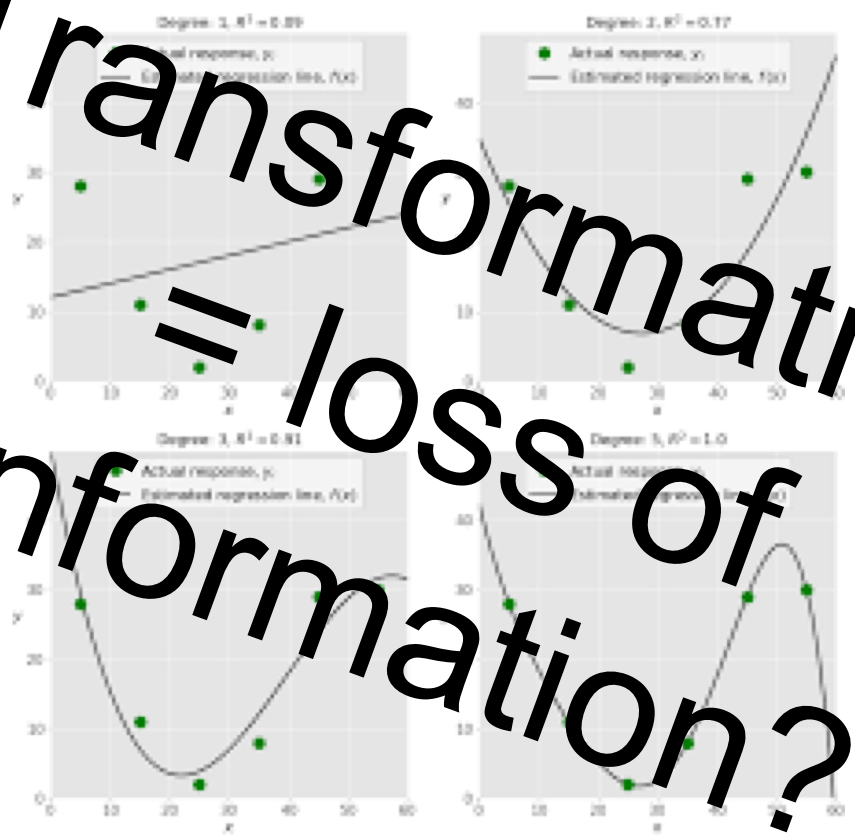


Image taken from realpython.com

Up-sell



Transformations ≡ loss of information?

Image taken from realpython.com

Consider...

Random Forest approach:

- powerful
- allows multicollinearity
- linear and non-linear effects within the same analysis[1]

Image taken from AsiaOne.com

1: Strobl, Malley & Tutz, 2009

Additional benefits:

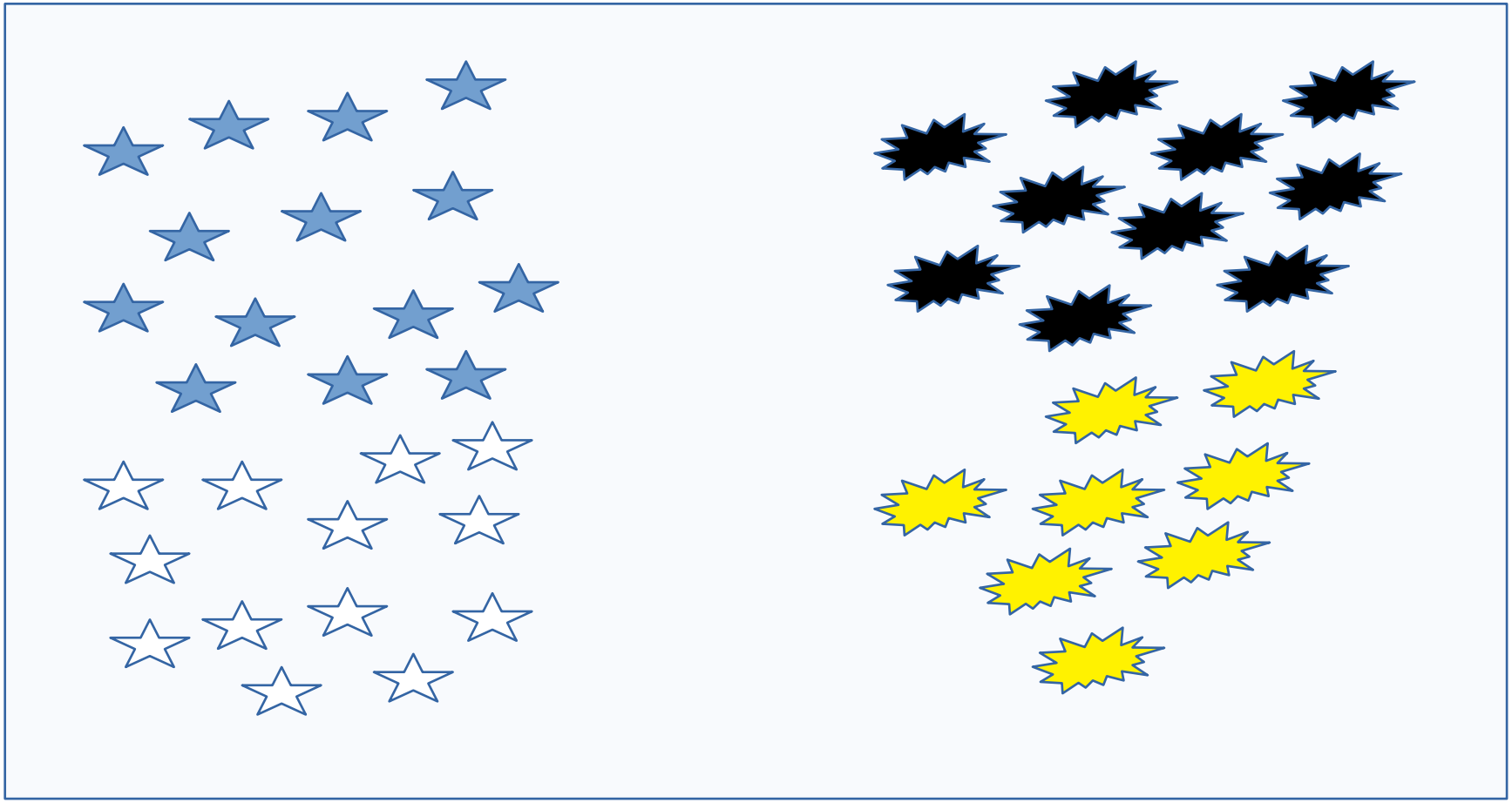
- Works with a small sample size
- Insensitive to order effects
  - Sampling process mitigates experimenter bias
- Can assist with variable reduction – if required[2]
  - More stable than stepwise regression[3]
  - Works with observed variables rather than latent variables[4]

2: Tagliamonte & Baayen, 2012 3: Strobl, Malley & Tutz, 2009 4: Breiman, 2001

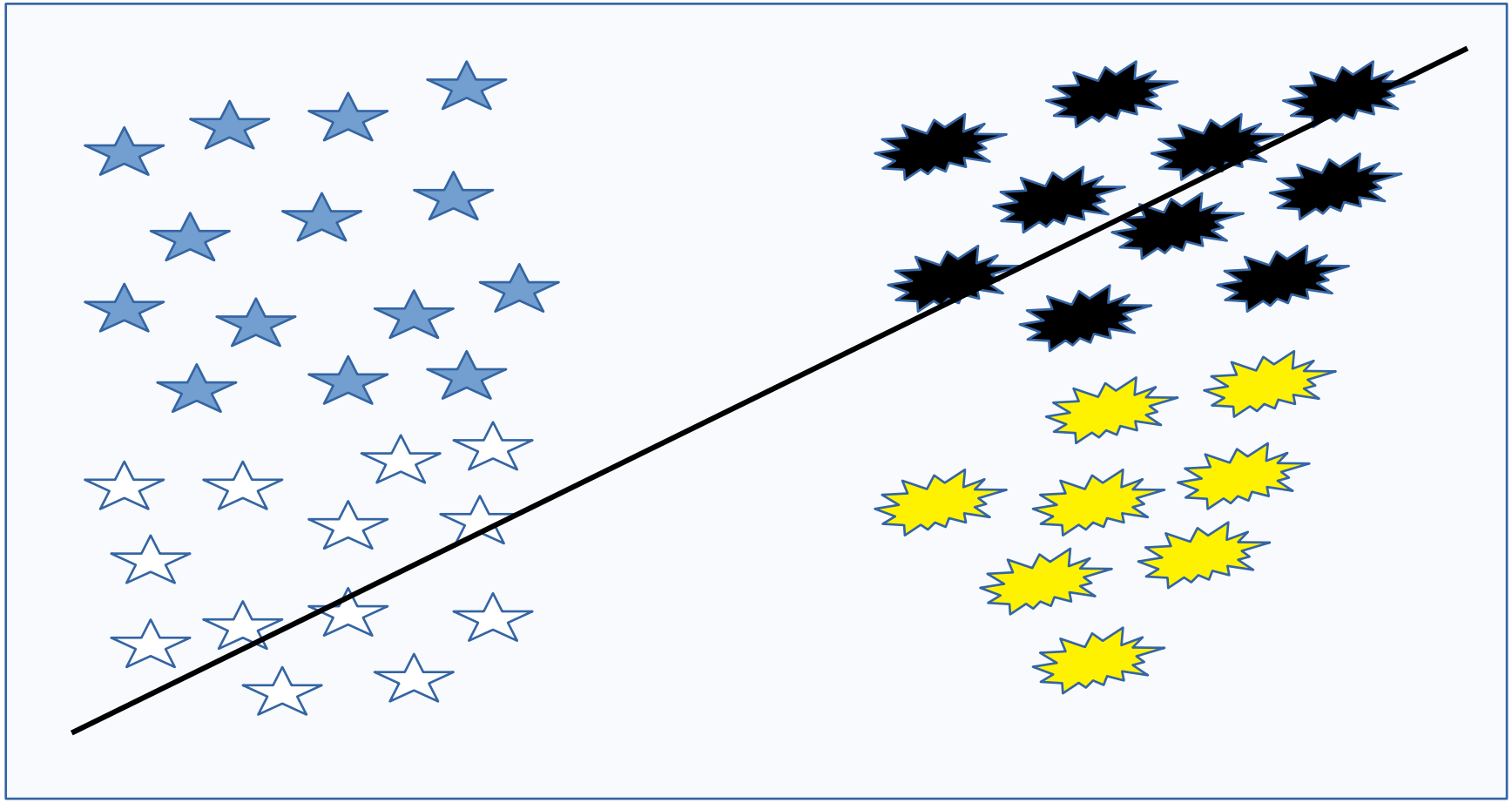
Health warnings:

- Black Box – a lot of stuff under the hood that you don't mess with unless you *really* know what you're doing!
- Not a replacement – more of a complimentary tool
- Data driven method – your data sample may not align with your theoretical expectations
- Variable selection takes a lot of time

How does it work? A simple example

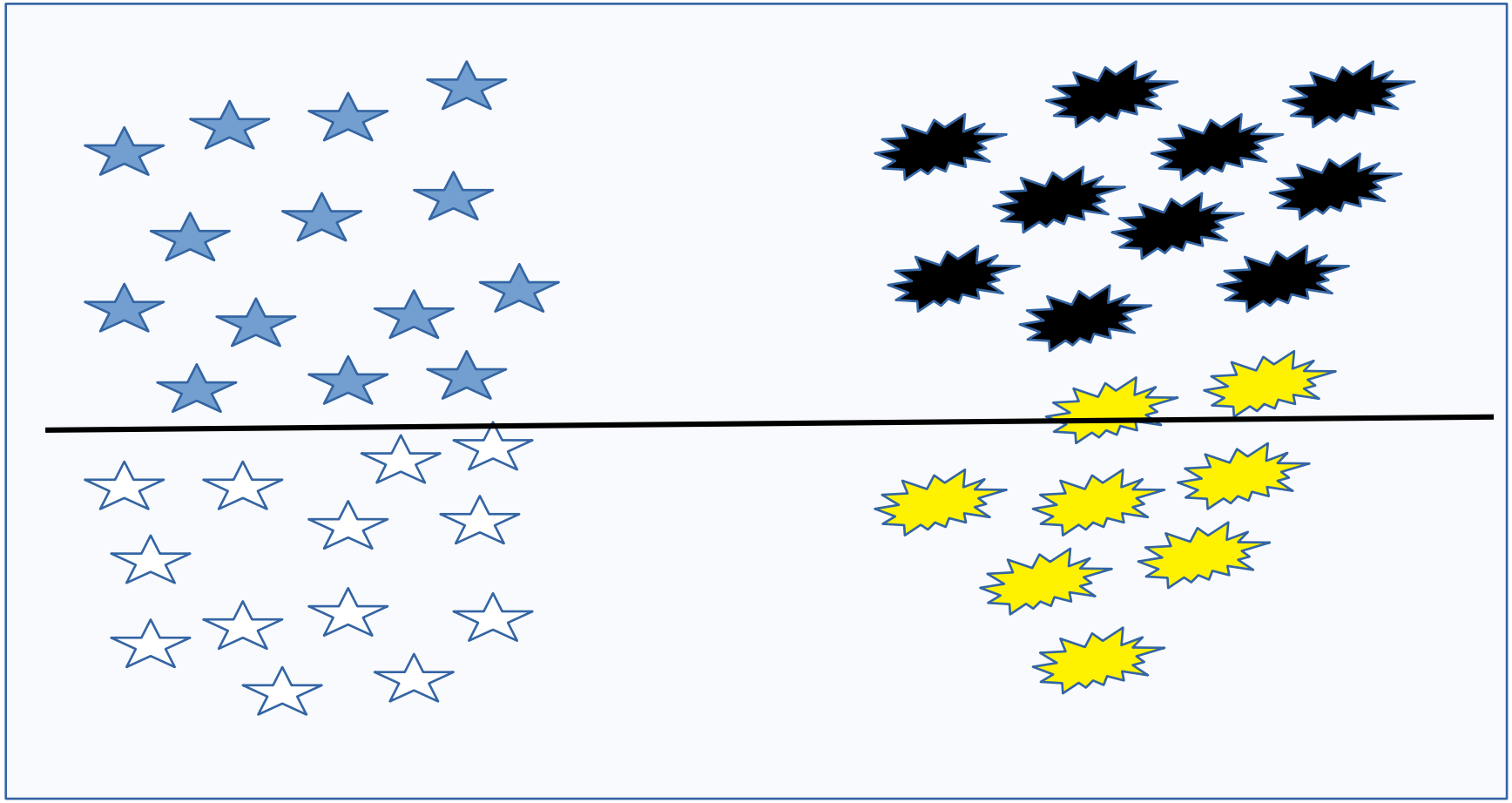


Random splits across the data set by one variable



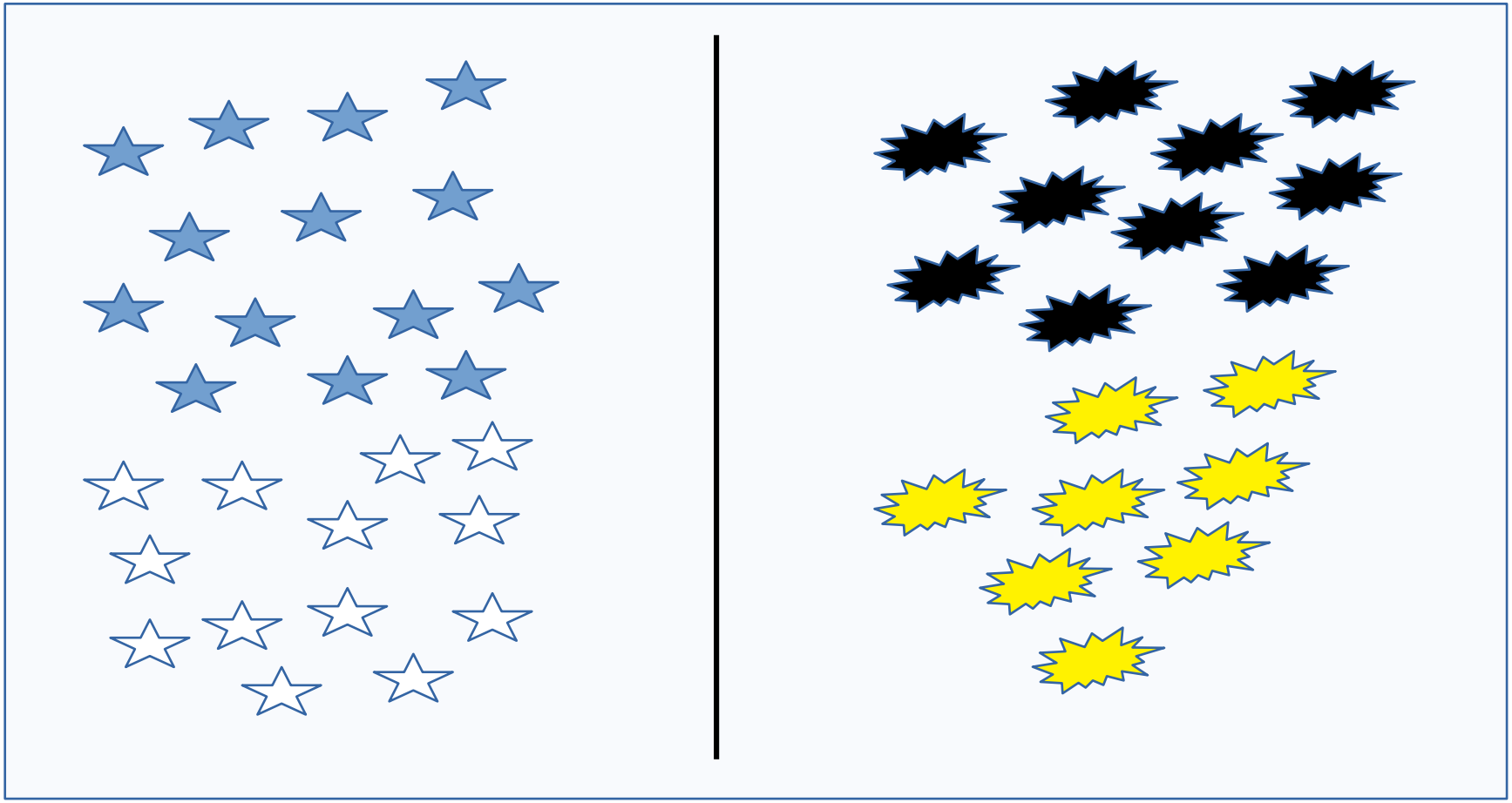
Lancaster University

Recursive partitioning occurs until...

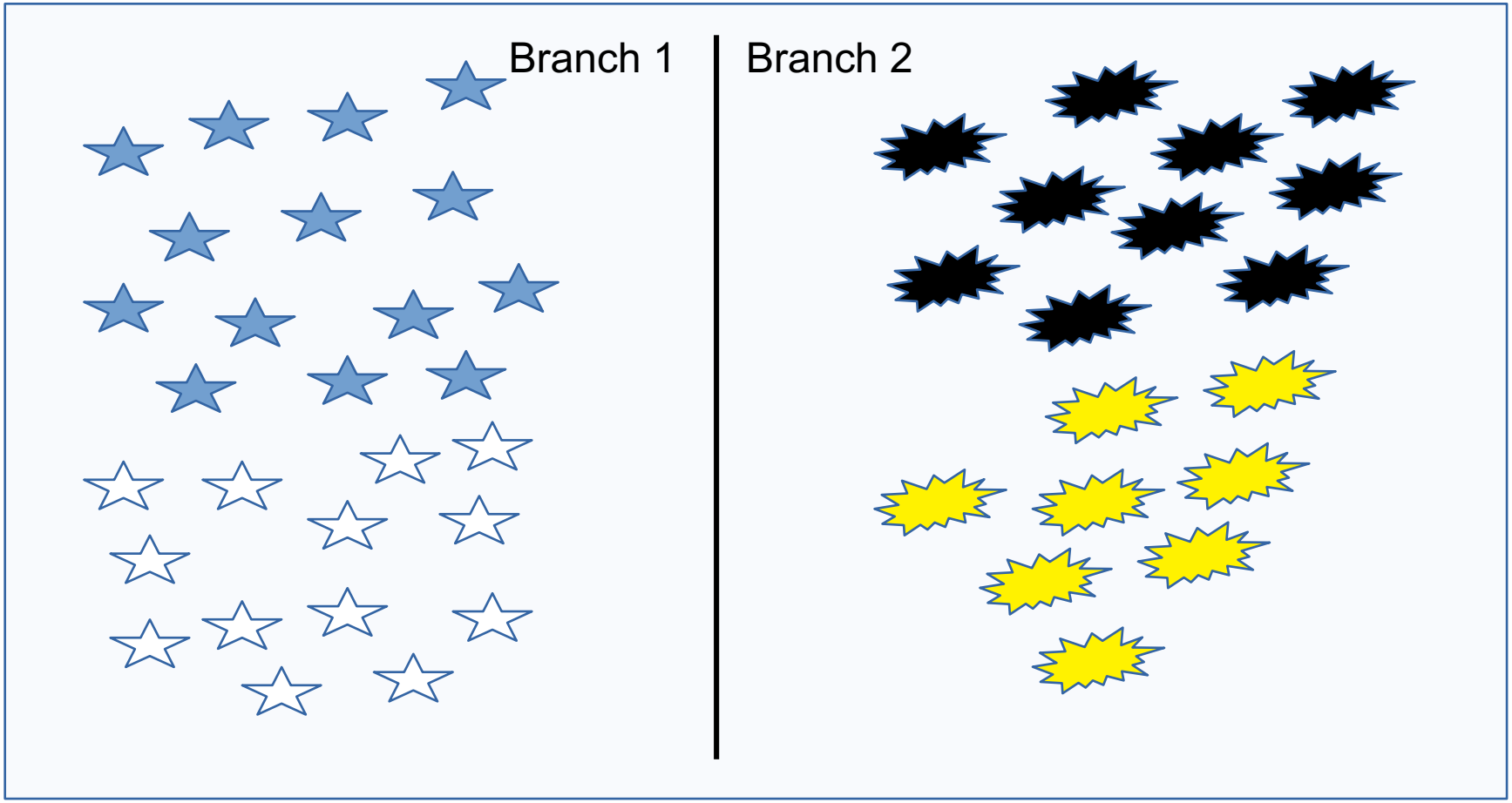


Lancaster University

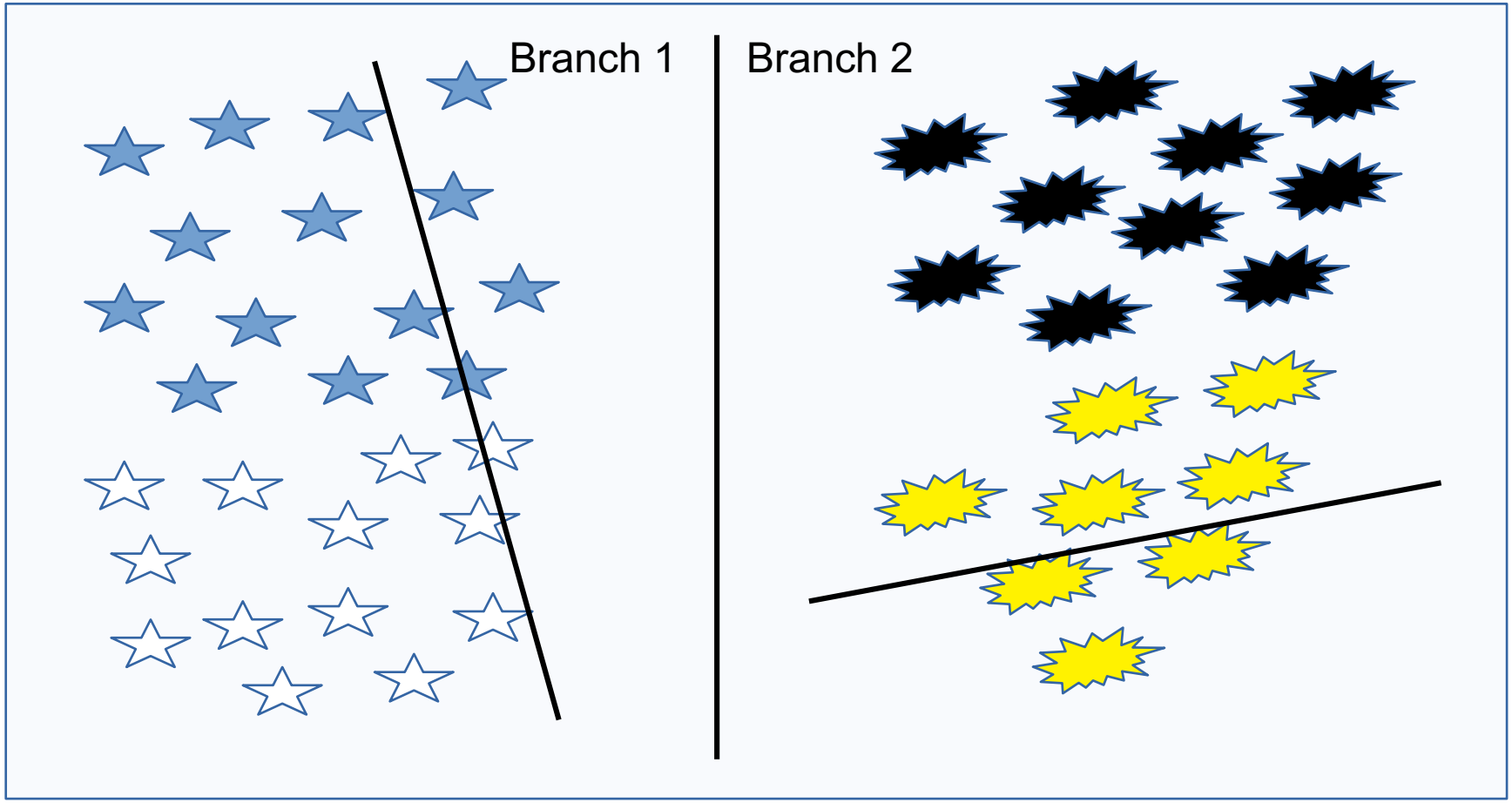
Impurity reduction



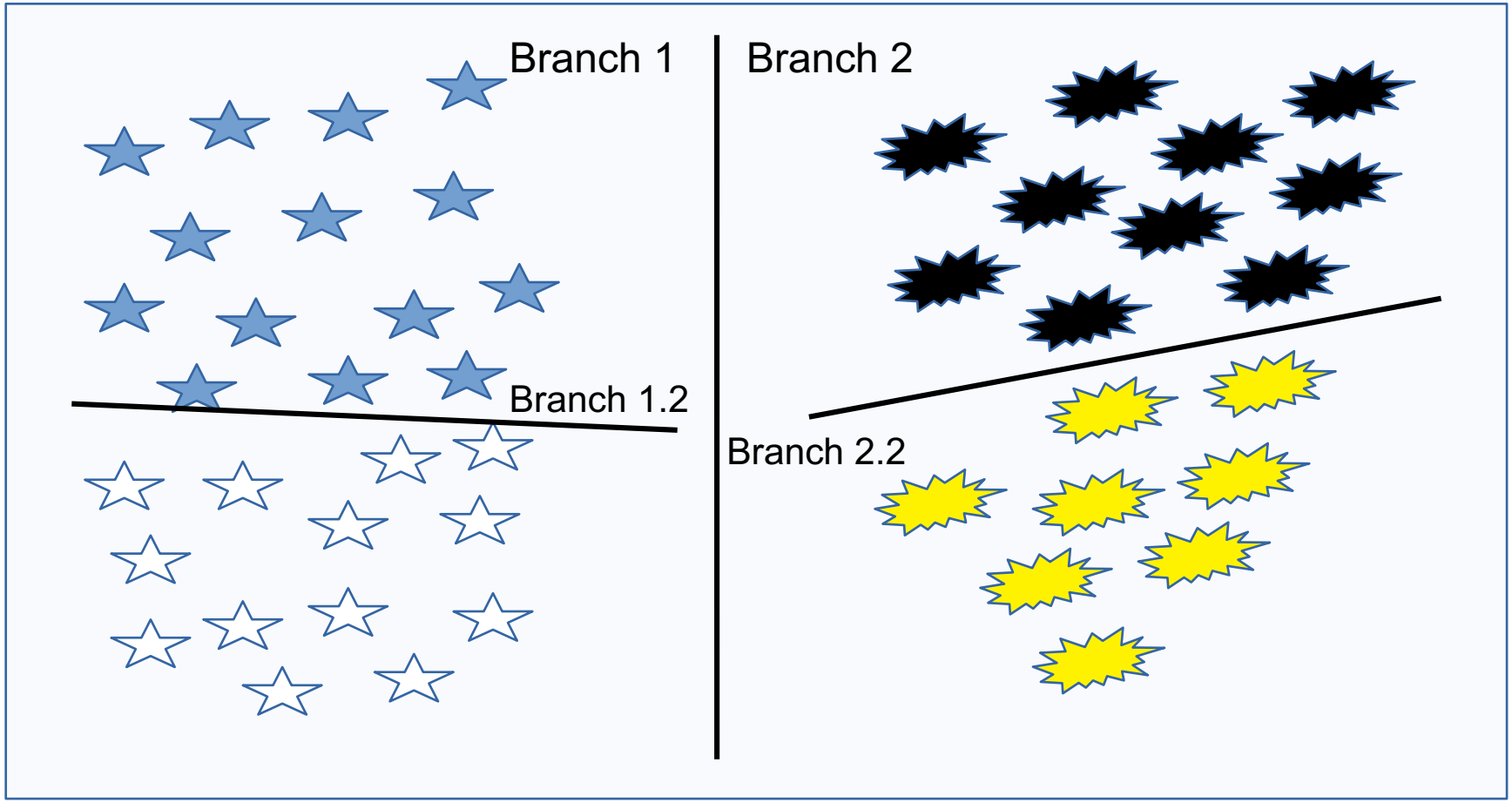
Binary split...



And repeat...by another variable



And repeat...



A slightly different example…. N = 100

2.31	2.31	2.31	2.31	2.31	2.31	6.48	6.48	6.48	6.48
2.31	2.31	2.31	2.31	2.31	2.31	6.48	6.48	6.48	6.48
2.31	2.31	2.31	2.31	2.31	2.31	6.54	6.55	6.55	6.55
2.31	2.31	2.31	2.31	2.31	2.31	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55

Split 1: < 6 - > 6

2.31	2.31	2.31	2.31	2.31	2.31	6.48	6.48	6.48	6.48
2.31	2.31	2.31	2.31	2.31	2.31	6.48	6.48	6.48	6.48
2.31	2.31	2.31	2.31	2.31	2.31	6.54	6.55	6.55	6.55
2.31	2.31	2.31	2.31	2.31	2.31	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55

Split 2: = 2.31 & > 2.31; = 6.48 & > 6.48

2.31	2.31	2.31	2.31	2.31	2.31	6.48	6.48	6.48	6.48
2.31	2.31	2.31	2.31	2.31	2.31	6.48	6.48	6.48	6.48
2.31	2.31	2.31	2.31	2.31	2.31	6.54	6.55	6.55	6.55
2.31	2.31	2.31	2.31	2.31	2.31	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55

Split 3: = 2.375 & = 2.379; = 6.48 & =6.55

2.31	2.31	2.31	2.31	2.31	2.31	6.48	6.48	6.48	6.48
2.31	2.31	2.31	2.31	2.31	2.31	6.48	6.48	6.48	6.48
2.31	2.31	2.31	2.31	2.31	2.31	6.54	6.55	6.55	6.55
2.31	2.31	2.31	2.31	2.31	2.31	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55
2.375	2.375	2.375	2.375	2.379	2.379	6.54	6.55	6.55	6.55

Lancaster University

Random Forest

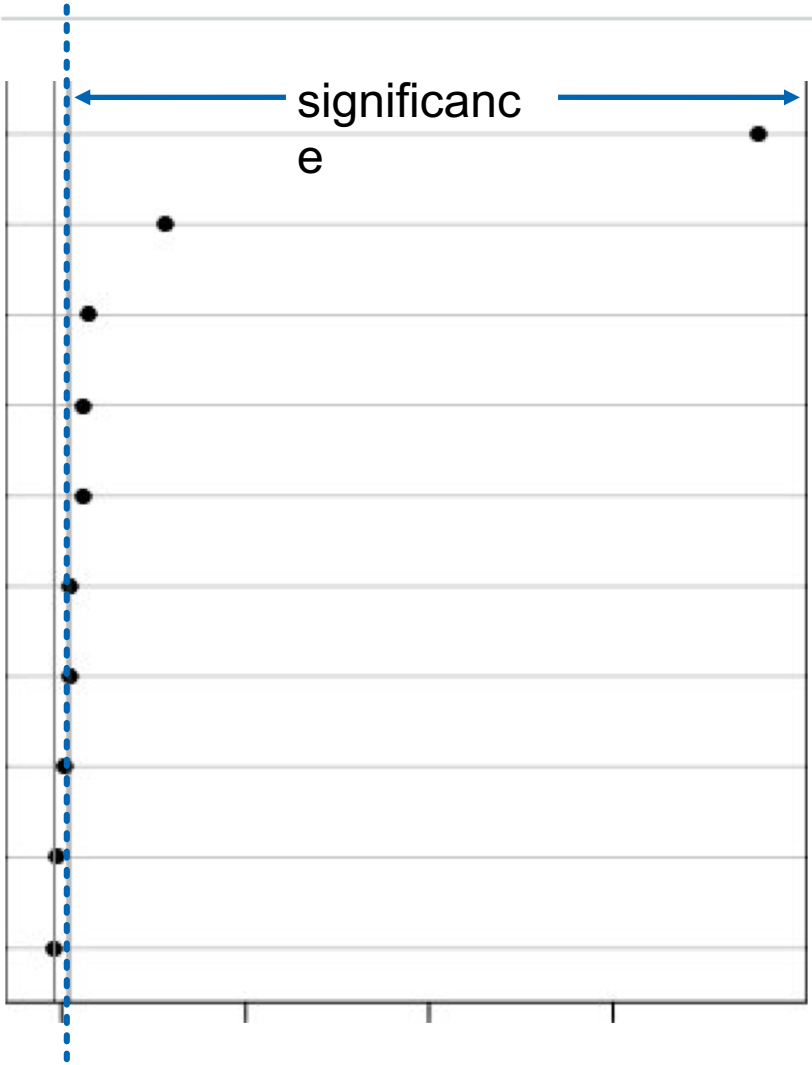
- One such tree is constructed many times with random sampling
- Random sampling without replacement
  - Across predictors
  - Different predictors as initial split
  - Across observations

Results

- Aggregation over all the trees: response variable with the most vote wins

- Variable importance needed though
  - Sampling means not all predictors are considered across all trees
  - Assessed by assigning new levels to variables and testing for an effect - a kind of sensitivity analysis but within and across variables
  - By product = cross validation: training and testing across full sample each time

Variable Importance Graphic



- Points to the right = significant variables
- Advice for variable selection is to *exclude* variables that are within the same range as *negative* variables
- Re-write formula with identified variables

Figure taken from Tagliamonte & Baayen 2012

Example plot: Main Effects & Interactions

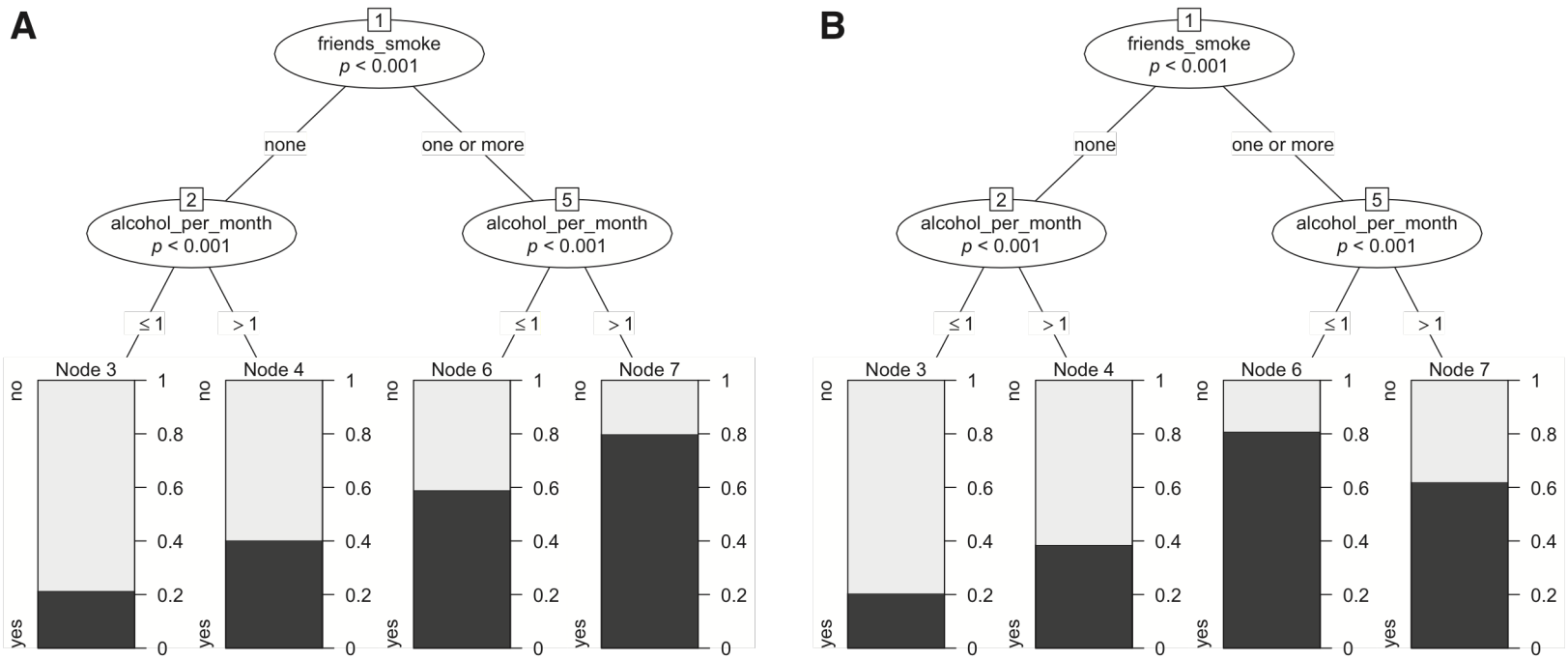


Figure 4. Classification trees based on variations of the smoking data with two main effects (Panel A) and interactions (Panel B). The tree depicted in Figure 1 that is based on the original data also represents an interaction.

Figure taken from Strobl, Malley & Tutz 2009

Conclusions

Random Forest offers

- Easy to implement method – formula just like regression
- Very flexible – small, non-linear sample – no problem
- Sampling method that validates itself in the process
- Variable selection routines if required
- Plots are quite intuitive to interpret

Starting resources

- ‘party’ package in R – vignettes and examples work well
- Tagliamonte & Baayen (2012) – comparison of regression, linear mixed effects and random forest with supplementary code
- Breiman (2001) – foundational concepts (pick and choose bits)
- Strobl, Malley & Tutz (2009) – update of Breiman for selection bias and other things (easier to read in its entirety than Breiman)

Example plot

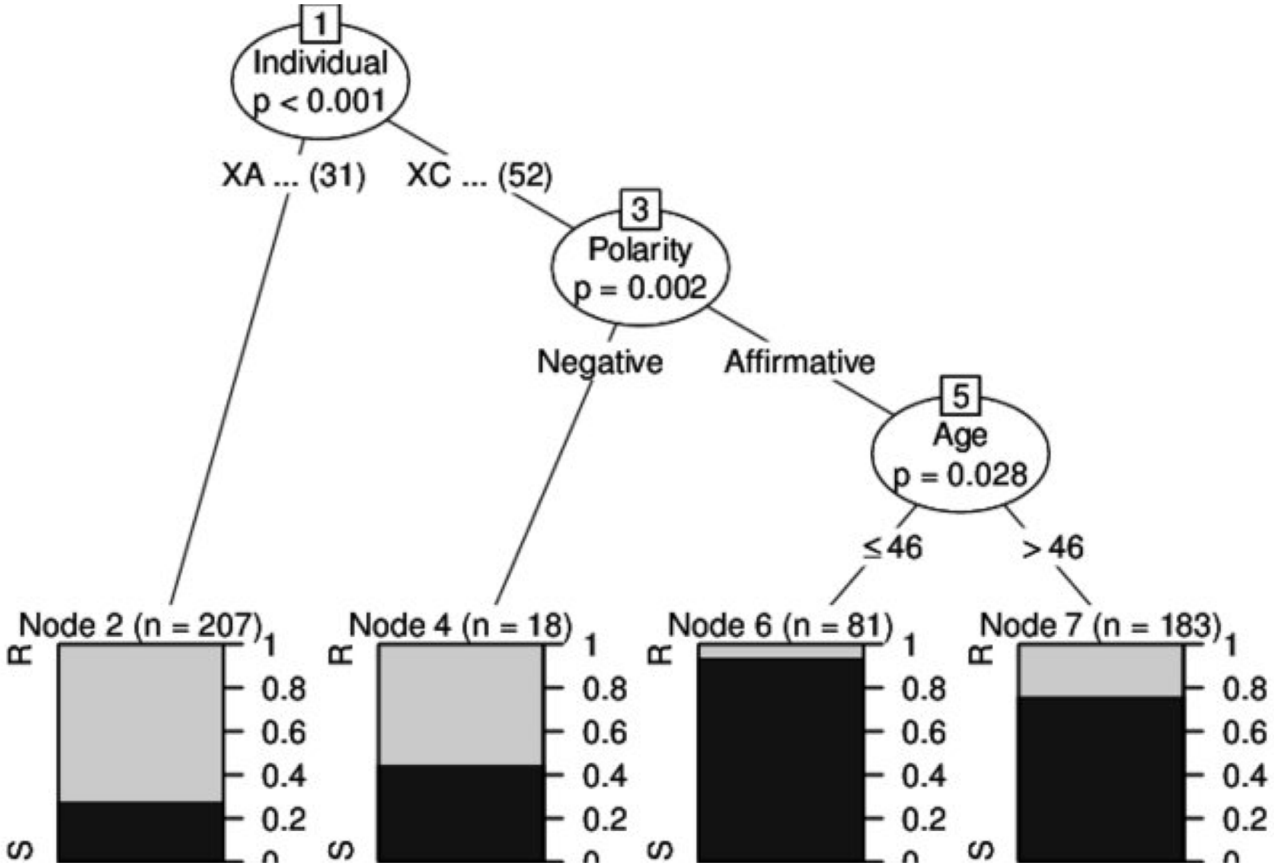


Figure taken from Tagliamonte & Baayen 2012