Seeing the forest for the trees

Marjolein Fokkema,

Leiden University

What's old

Regression models used for:

- *Prediction*
- *Inference*

Computer cluster (1920)

What's new

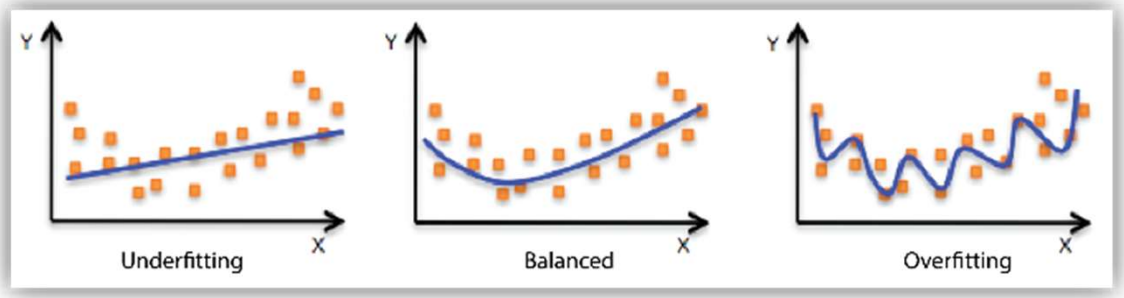
- More computational power
- Bigger data (both *N* and *p*)

-> wider use and implementation of flexible algorithms:

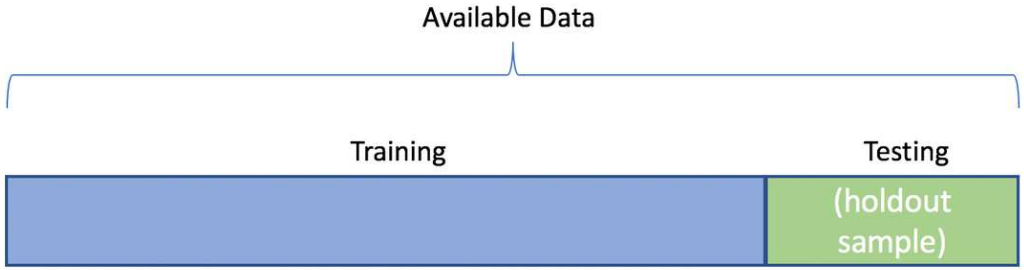
  - Neural nets (1958, Rosenblatt)
  - Decision trees (1959, Belson / 1980, Kass)
  - Support vector machines (1963, Vapnik & Chervonenkis)
  - Smoothing splines (1946, Schoenberg)
  - k Nearest neighbors (1951, Fix & Hodges)
  - Penalized regression (1970, Hoerl & Kennard)

What's new

- Increased flexibility:



- Need built-in overfitting control

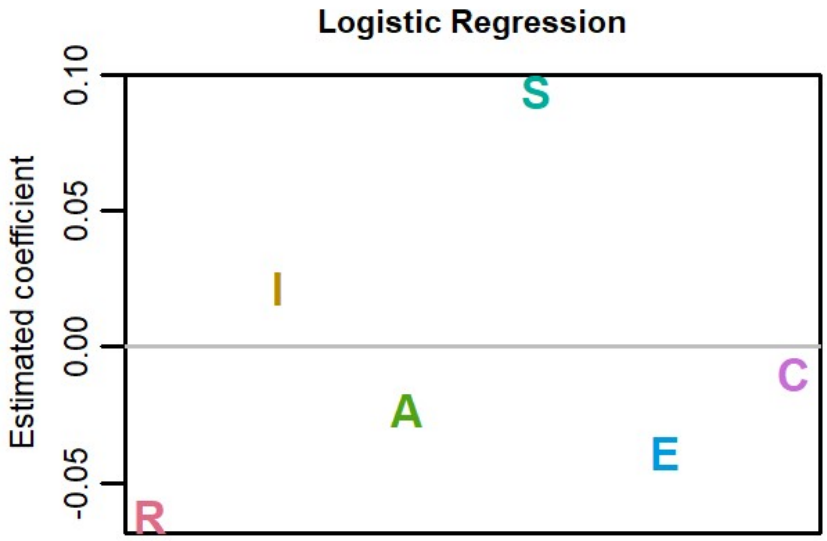


- Increased focus on prediction of new cases / generalizability

- Not new to psychometrics, e.g.:
  - Larson (1931), Mosier (1951): Cross validation & generalizability
  - Darlington (1978): Reduced variance regression
  - Gifi (1981) methods: Focus on minimizing loss functions & assume no statistical model
  - Mixed-effects models

Dataset

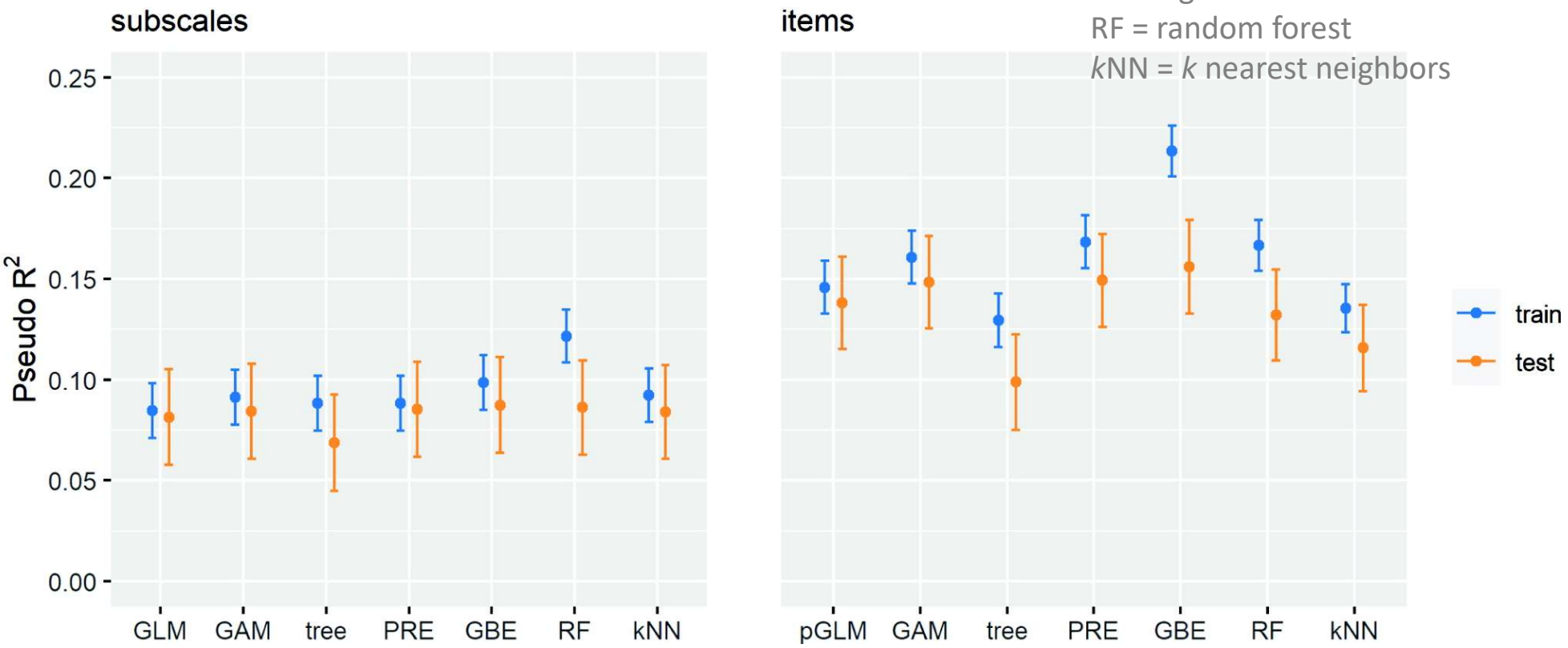
- Open Psychometrics Project (2015 - 2018)
  - N = 55,593 (75% training; 25% test)
- Outcome:
  - Took psychology as a major at university: Yes (19.4%) vs. No
- Predictors:
  - RIASEC vocational preferences scales
  - Realistic, Investigative, Artistic, Social, Enterprising, Conventional
- For all methods, parameter settings tuned using 10-fold CV on training data



- Realistic, Investigative, Artistic, Social, Enterprising, Conventional

Results

(p)GLM = (penalized) logistic regression
GAM = generalized additive model with smoothing splines
PRE = prediction rule ensemble
GBE = gradient boosted tree ensemble
RF = random forest
*k*NN = *k* nearest neighbors



Intermediate conclusion

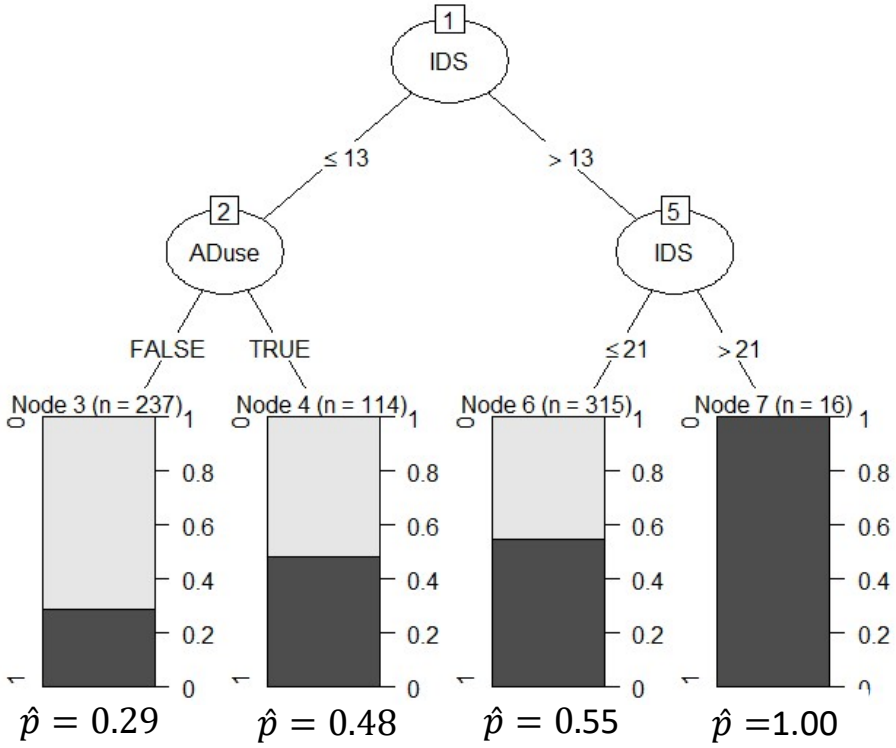
- Simple methods capture most of the signal
- Sophisticated ML methods can improve, but often marginally
- Gains may be swamped by practical aspects (Hand, 2006; Efron, 2020)
  - Measurement / labelling errors
  - Population drift
  - Need for interpretability
  - Cost of information
- Fokkema, Iliescu (in press) European Journal of Psychological Assessment

-> Can have simple, interpretable ML model with near-optimal accuracy?

Dataset: Predicting depression

- Respondents with current depressive disorder (N = 682)
- Response: Depression diagnosis (at two-year follow-up)
- 20 possible predictors (at baseline):
  - Age
  - Gender
  - Education level
  - Anxiety disorder
  - Symptom severity
  - Treatment
  - …

Single decision tree for predicting depression



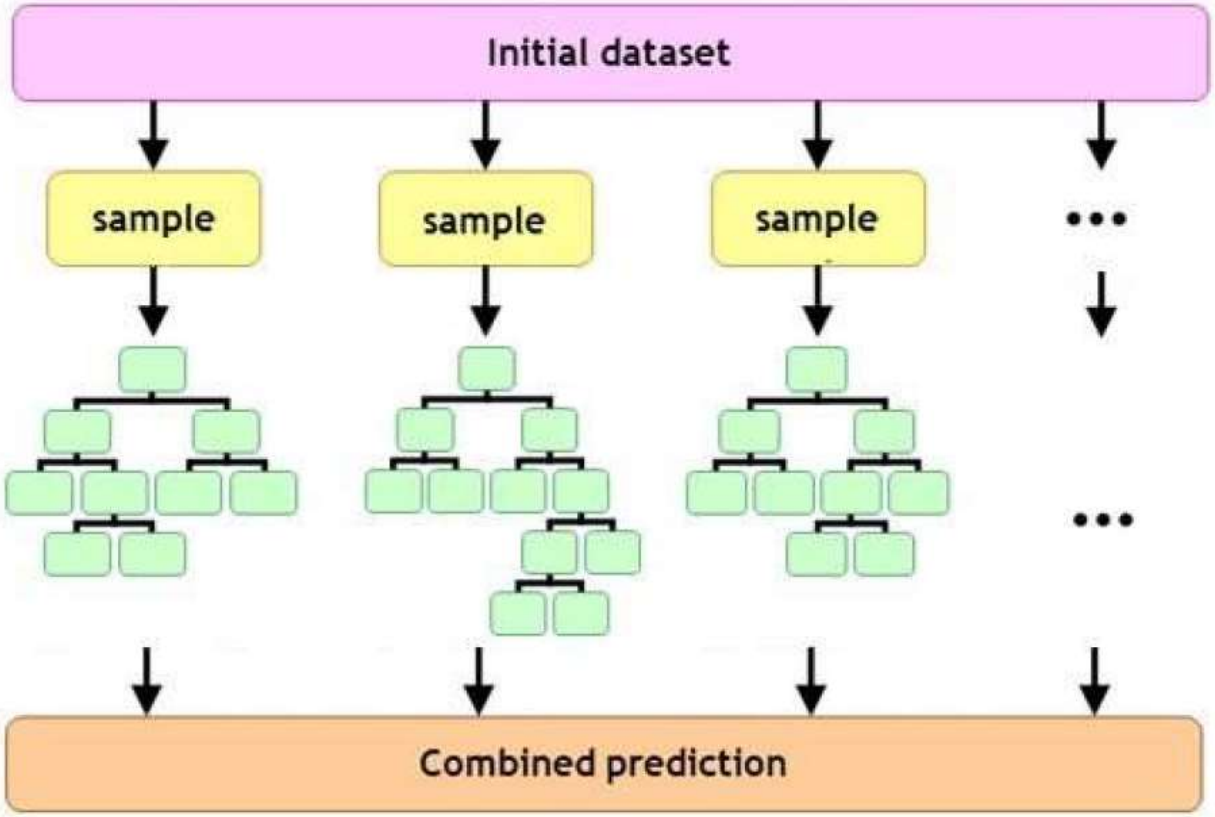
Decision trees

Good: Easy to interpret and apply

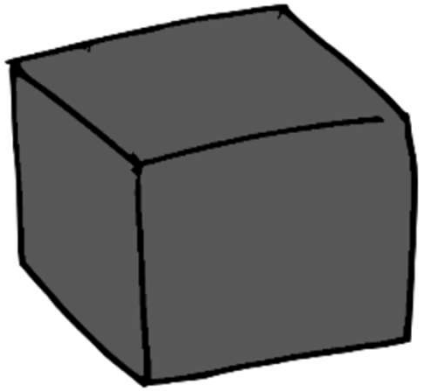
Bad: Not most accurate method

Ugly: Unstable

Decision tree ensembles



- Bagging
- Random forests
- Gradient boosting
- .....



RuleFit algorithm (Friedman & Popescu, 2008)

1) Take subsamples from training data
2) Grow tree on each sample
3) Extract (very large) initial ensemble of rules:
   - Include every node from every tree as a rule and
   - Include original predictor variables as linear terms
4) Select smaller final ensemble by sparse regression on training data:
   - Lasso regression

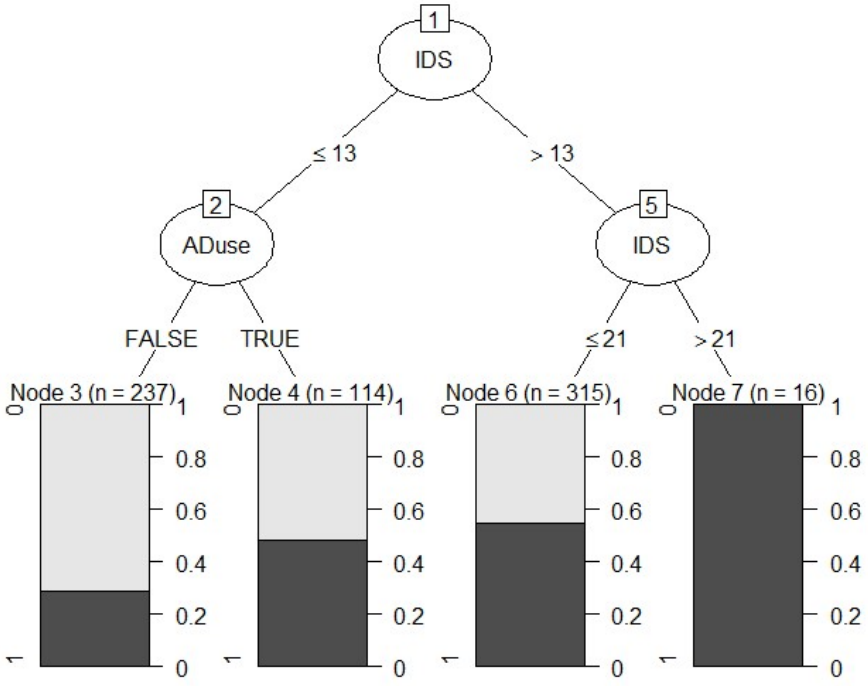
R package pre (Fokkema & Christoffersen)

Implements and extends RuleFit algorithm:

+ multivariate, multinomial, count, survival responses
+ use of unbiased tree algorithm
+ (non-)negativity constraints
+ include confirmatory rules
+ relaxed lasso
…

From trees to rules

$r_2(\mathbf{x}) = I(IDS \leq 13)$
$r_3(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = FALSE)$
$r_4(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = TRUE)$
~~$r_5(\mathbf{x}) = I(IDS > 13)$~~
$r_6(\mathbf{x}) = I(IDS > 13) \cdot I(IDS \leq 21)$
$r_7(\mathbf{x}) = I(IDS > 13) \cdot I(IDS > 21)$



From trees to rules

$r_2(\mathbf{x}) = I(IDS \leq 13)$
$r_3(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = FALSE)$
$r_4(\mathbf{x}) = I(IDS \leq 13) \cdot I(ADuse = TRUE)$
$r_6(\mathbf{x}) = I(IDS > 13) \cdot I(IDS \leq 21)$
$r_7(\mathbf{x}) = I(IDS > 13) \cdot I(IDS > 21)$
…
$l_1(\mathbf{x}) = IDS$
$l_2(\mathbf{x}) = ADuse$
…

$$F(\mathbf{x}) = \hat{\alpha}_0 + \sum_{m=1}^{M} \alpha_m f_m(\mathbf{x})$$

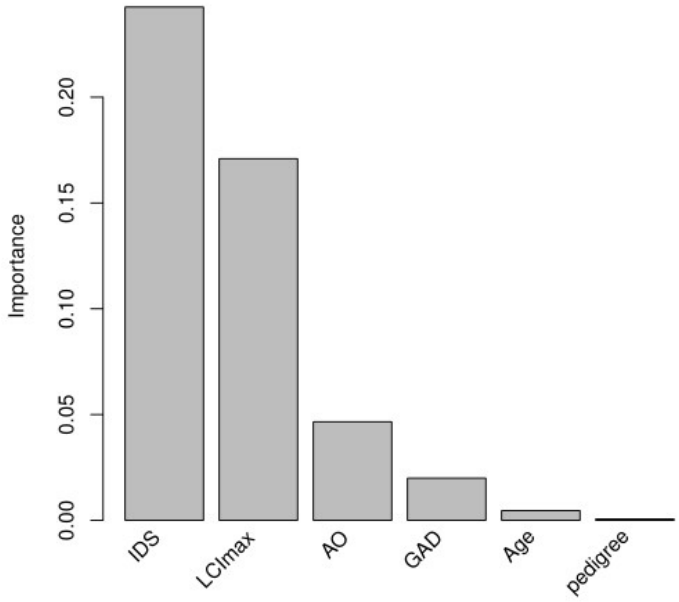
IDS	ADuse	…	r_2	r_3	r_4	r_6	r_7	…
5	FALSE	…	1	1	0	0	0	…
15	FALSE	…	0	0	0	1	0	…
18	TRUE	…	0	0	0	1	0	…
25	TRUE	…	0	0	0	0	1	…
…	…	…	…	…	…	…	…	…

PRE for predicting depression (Fokkema & Strobl, 2020)

Term	Description	Coeffient	SD	Importance
(Intercept)	1	-0.221	0.000	0.000
rule3	IDS > 10 & LCImax > 0.2632	0.224	0.494	0.111
rule27	IDS > 13 & LCImax > 0.3621	0.213	0.477	0.102
rule84	IDS <= 16 & AO > 17	-0.175	0.489	0.086
rule67	IDS > 10 & LCImax > 0.3276	0.140	0.500	0.070
rule51	LCImax > 0.26 & IDS > 9	0.122	0.487	0.059
rule24	IDS <= 16 & GAD %in% c(”Negative”)	-0.080	0.499	0.040
rule110	IDS > 10 & Age > 22	0.020	0.459	0.009
rule125	IDS <= 17 & AO > 13	-0.015	0.496	0.007
rule108	IDS > 14 & pedigree %in% c(”Yes”)	0.002	0.478	0.001

Table 1: Predicting chronic depression - Default settings.

Variable importances (Fokkema & Strobl, 2020)

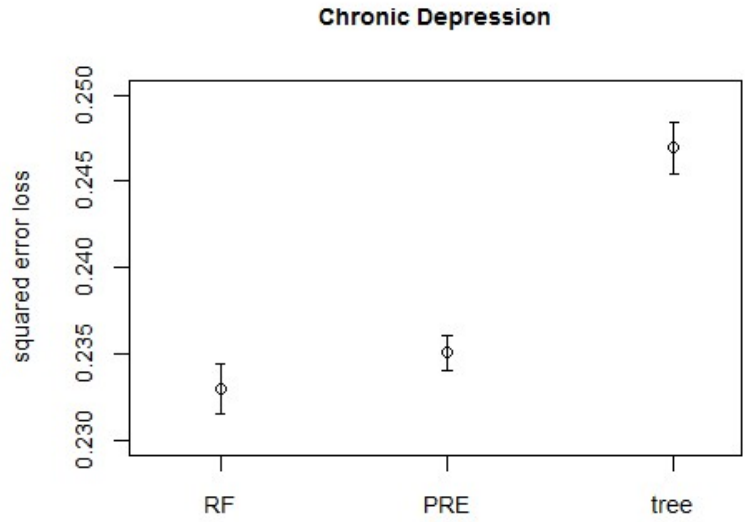


Term	Description	Coeffient	SD	Importance
(Intercept)	1	-0.221	0.000	0.000
rule3	IDS > 10 & LCImax > 0.2632	0.224	0.494	0.111
rule27	IDS > 13 & LCImax > 0.3621	0.213	0.477	0.102
rule84	IDS <= 16 & AO > 17	-0.175	0.489	0.086
rule67	IDS > 10 & LCImax > 0.3276	0.140	0.500	0.070
rule51	LCImax > 0.26 & IDS > 9	0.122	0.487	0.059
rule24	IDS <= 16 & GAD %in% c("Negative")	-0.080	0.499	0.040
rule110	IDS > 10 & Age > 22	0.020	0.459	0.009
rule125	IDS <= 17 & AO > 13	-0.015	0.496	0.007
rule108	IDS > 14 & pedigree %in% c("Yes")	0.002	0.478	0.001

Table 1: Predicting chronic depression - Default settings.

Resolution

Fokkema & Strobl (2020):



Fokkema (2020):

Predictive accuracy: random forest > pre > RuleFit > linear lasso > single tree

Complexity: random forest > linear lasso > RuleFit > pre > single tree

Contributions & outlook

pre provides predictive accuracy close to tree ensembles

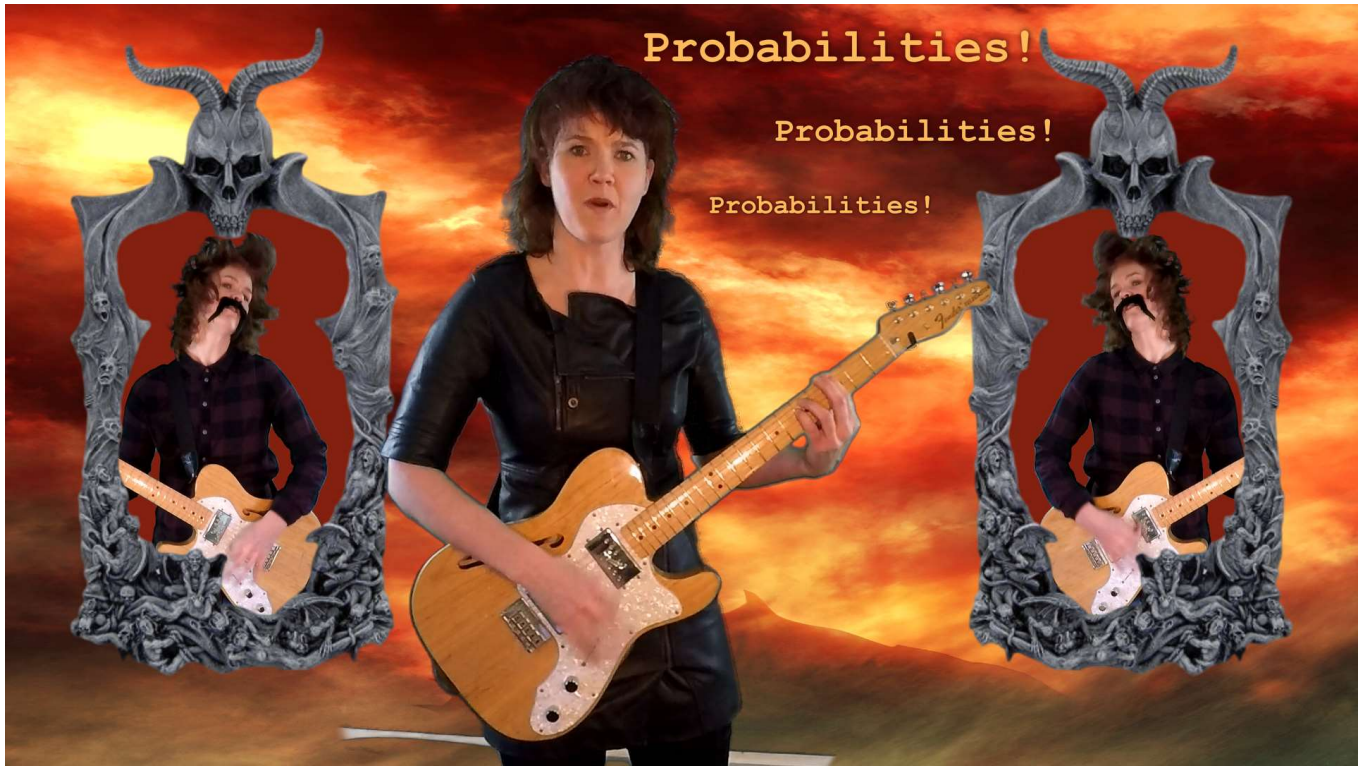
- More interpretable
- Uses less variables for prediction

Improving trade-off and control of complexity & accuracy

- Relaxed lasso (implemented)
- Example-generating approaches (Markovitch & Fokkema, 2021)

Inference and uncertainty quantification!

- Bayesian rule ensembles
- Causal rule ensembles



Mary-Jo & the Support Vector Machines

https://www.youtube.com/channel/UCP2uiCpatuCs4FceZZnJdTg

Fokkema, M., Iliescu, D., Greiff, S., & Ziegler, M. (in press). Machine learning and prediction in psychological assessment. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000714

Iliescu, D., Greiff, S., Ziegler, M., & Fokkema, M. (in press). Artificial intelligence, machine learning, and other demons. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000713

Prediction rule ensembles

Fokkema, M. & Christoffersen, B. (2019). **pre**: Prediction Rule Ensembles. **R** package.

Tutorials: https://CRAN.R-project.org/package=pre and https://github.com/marjoleinF/pre

Fokkema, M. (2020). Fitting prediction rule ensembles with **R** package **pre**. *Journal of Statistical Software, 92*(12), 1-30.

Fokkema, M. & Strobl, C. (2020). Fitting prediction rule ensembles to psychological research data: An introduction and tutorial. *Psychological Methods, 25*(5), 636–652.

Markovitch, B., & Fokkema, M. (2021). Improved prediction rule ensembling through model-based data generation. *arXiv preprint arXiv:2109.13672*.