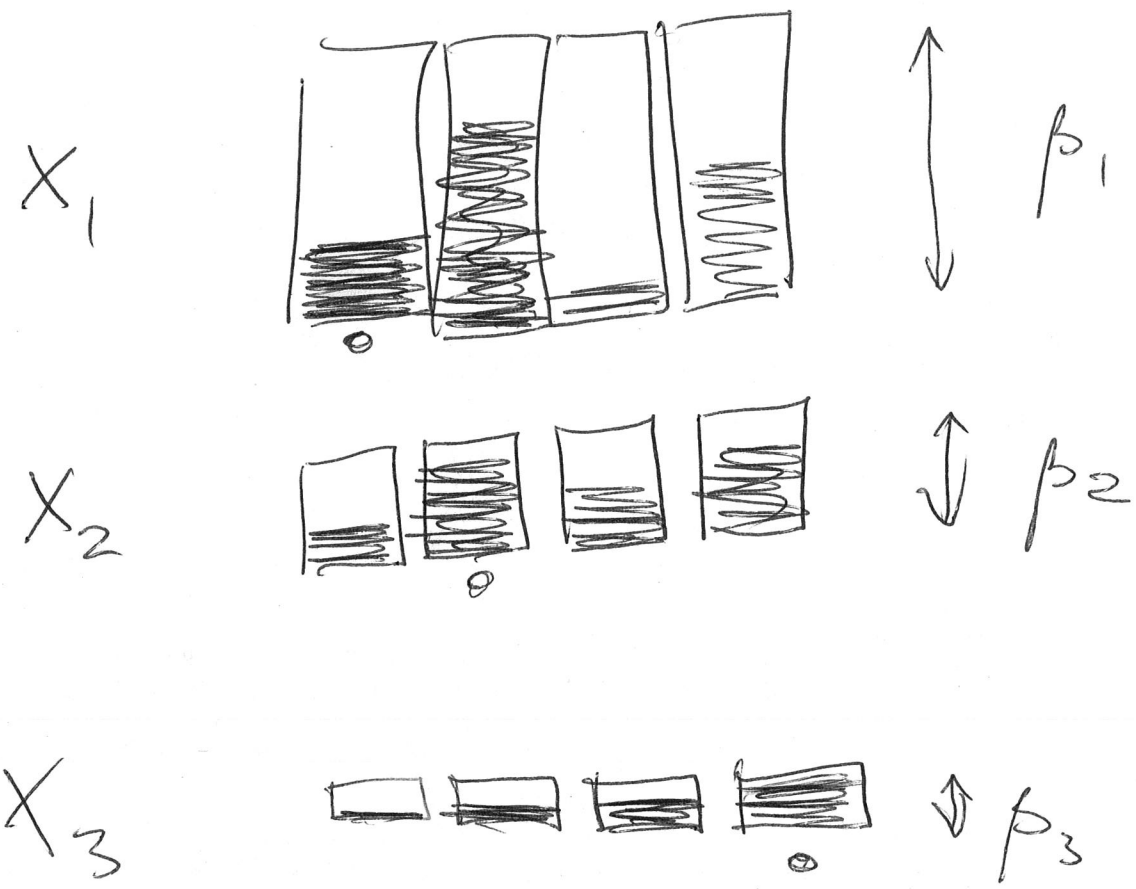
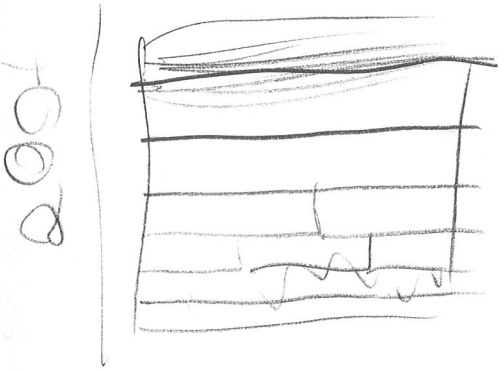
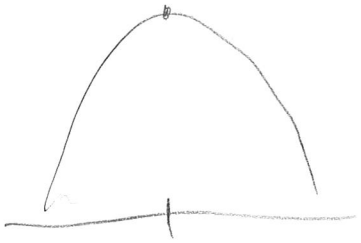

X_1
β_1
X_2
β_2
X_3
β_3
label
coeff.
X_1
X_2
X_3
$|X_1 \cap X_2 \cap X_3| = \frac{1}{64} \cdot n$

Pred	Obs	dop	Obs	rough_cor_cf
1	1	rough. cf cf	1	(0.3 0.7 | 1 0)
10	1	rough_cor cf		
1	2			
10	2	cf		



$$p_i = (p_{i1}, \ldots, p_{iq}) \text{ multi}$$



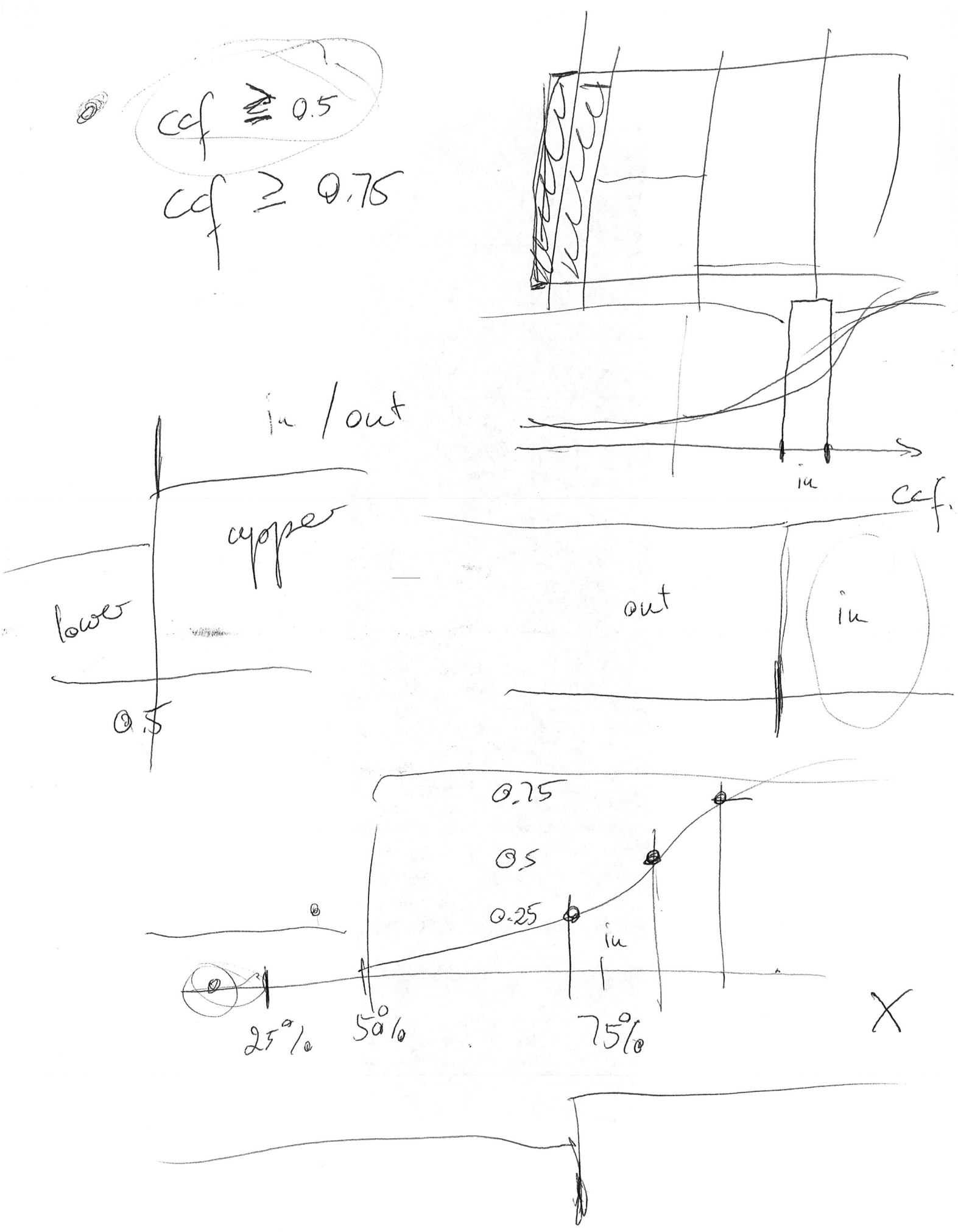
$p(1-p)$

$$\exp\left(-\sum_{k=1}^{q} p_{ik} \log p_{ik}\right)$$

Shannon entropy

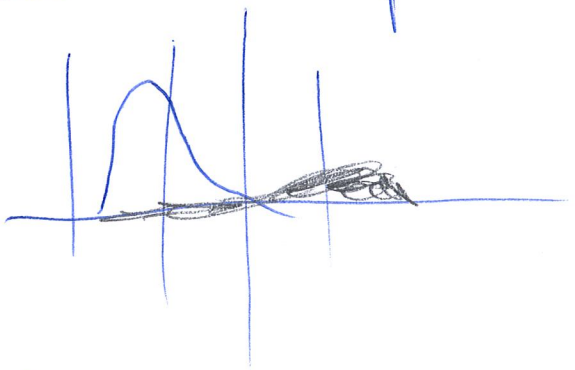
$$\left(1 - \sum_{k=1}^{q} p_{ik}^2\right) \in [0,1]$$

$$\sum_{k=1}^{q} p_{ik} = 1$$


cf ≧ 0.5
cf ≧ 0.75
in / out
upper
lower
0.5
in
cf.
out
in
0.75
0.5
0.25
in
25%
50%
75%
X

~~Same~~ Equidsin method

→ skew data will be problematic



Quantile method is better

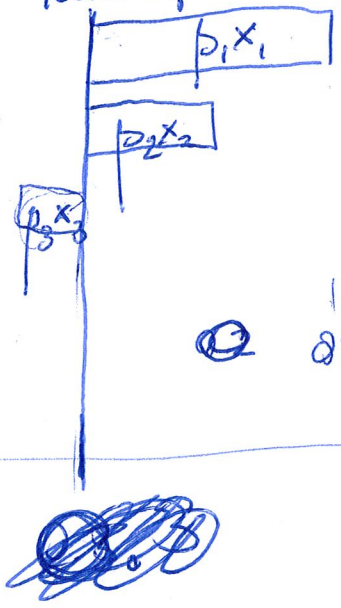
binning + ~~lasso~~ ridge + regression

$$y_{ij} = \sum b_j x_{ij} + \text{penalty} + \varepsilon$$

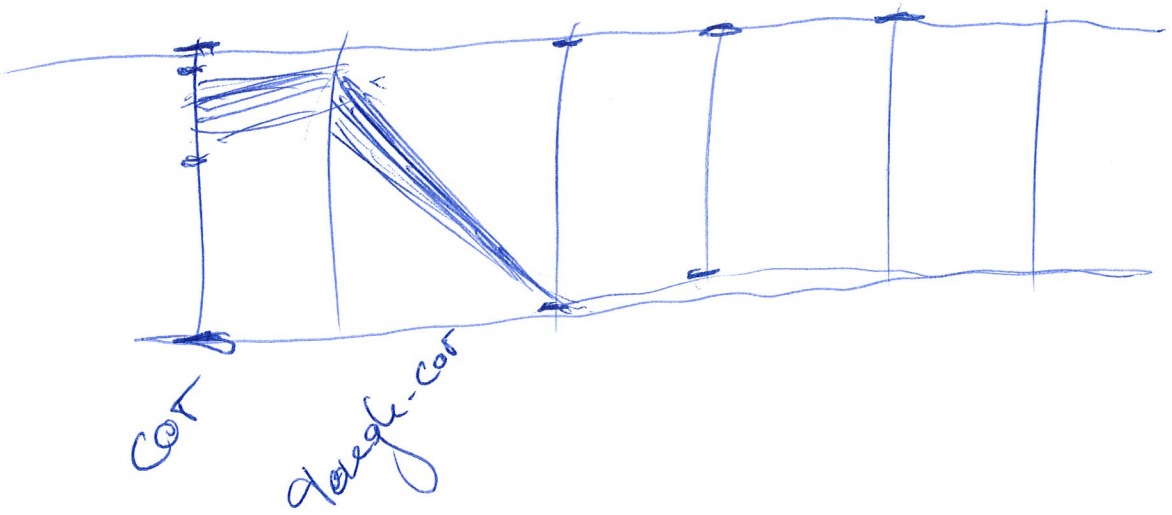
case i:

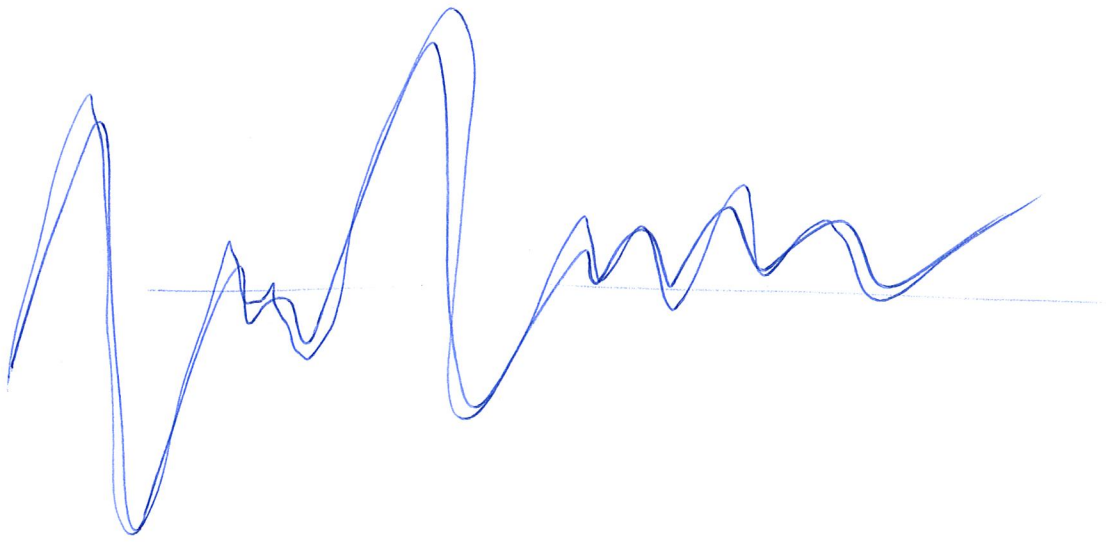
$b_j x_{ij}$ for each $j = 1, \dots p$

intercept is 0.3



col = rgb(1, 1, 1, alpha = 0.1)





ASA Data Science Journal

Visual Diagnostics of a model explainer

at the example of LIME

main objective of model explainer:

- understand and explain model performance

LIME does....

Conceptually: models at two levels:

explainer model — "simple"

original "black box" model — "complicated"

Usually: model predictions, maybe with ground truth

Type I error

Type II error

model is wrong

Explanation is also a prediction —

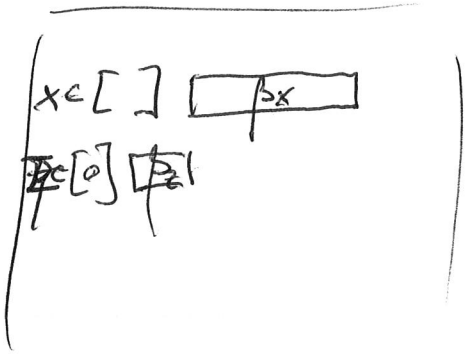
how reliable is that explanation?

Explains model has very low R^2 generally
- probably due to binning

"Local" explanations are not local, but are driven by the (global) marginal distribution of covariates

Describe LIME, including details on binning and linear regression in binned features

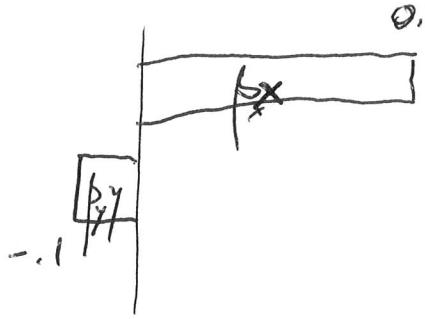
Motivation



0.3

$x = 8$

$y = 0.5$



$\mu = 0.5$

.sty
style file

Rmarkdown:

@bibtexlag — Author (2010)

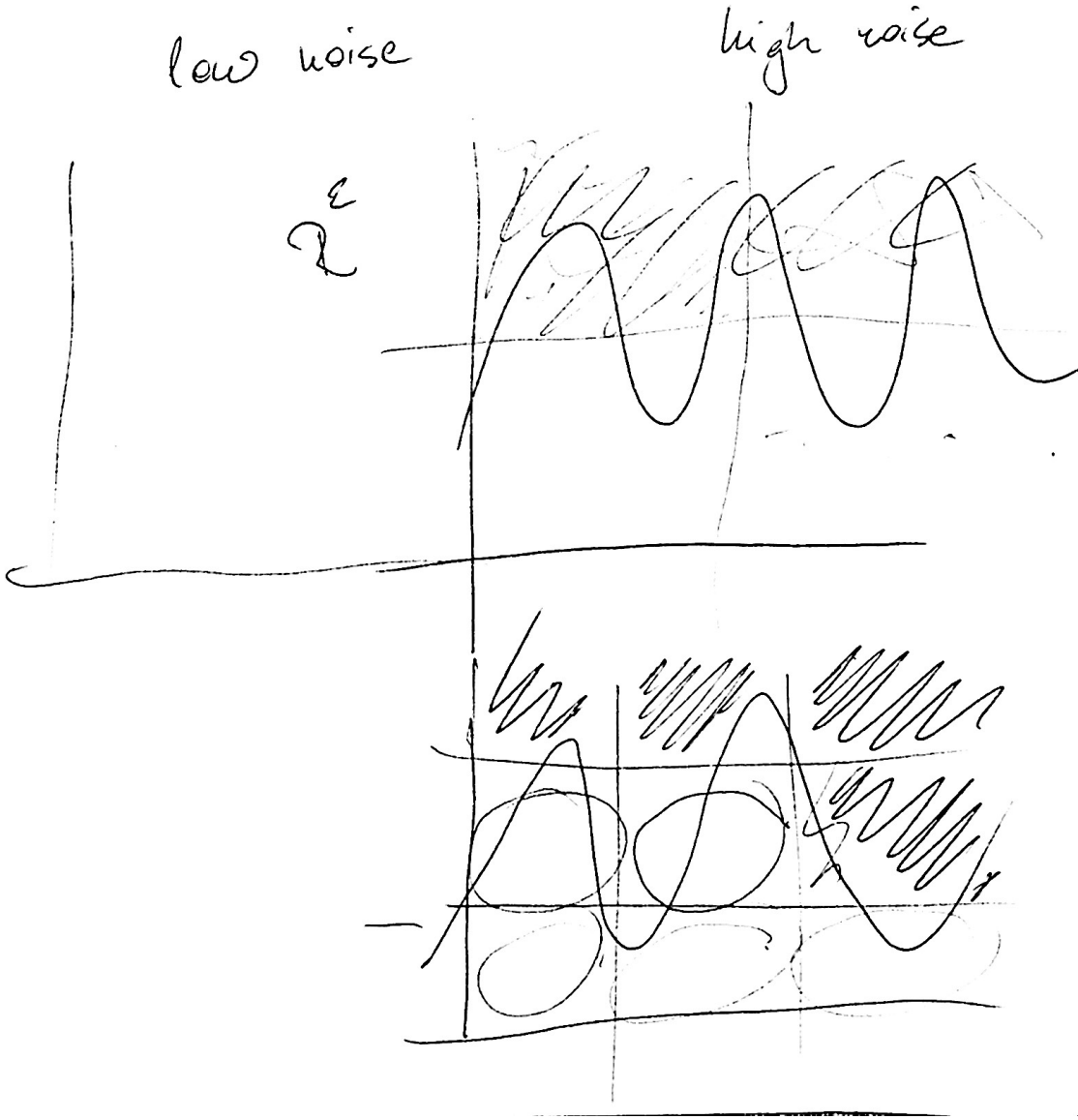
[@bibtexlag] — (Author 2010)

[@ref1; @ref2] — (Author1 2010, Author2 et al 2013)

yaml:

bibliography: name.bib


x_2
x_1


low noise
high noise
$\hat{x}^\varepsilon$

Local (interpretable) model agnostic explanations

4^2 is low number

~~simple~~

4^p might not be very low

Expectations for explanations

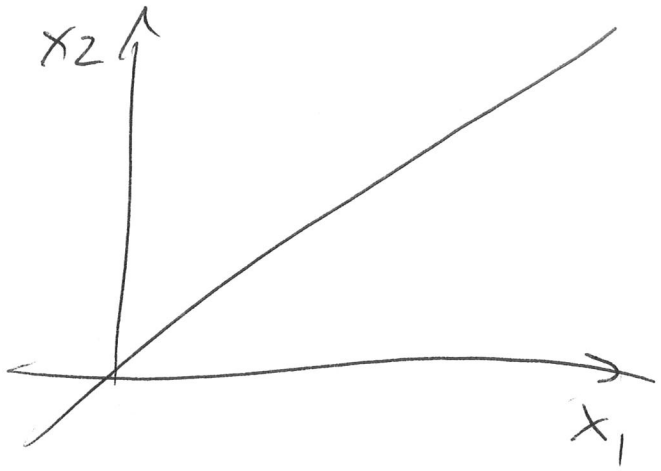
- data driven by relevant features
- "explain": difference in coefficients

deterministic/ non-deterministic

Random forest variable importance for identify relevant features

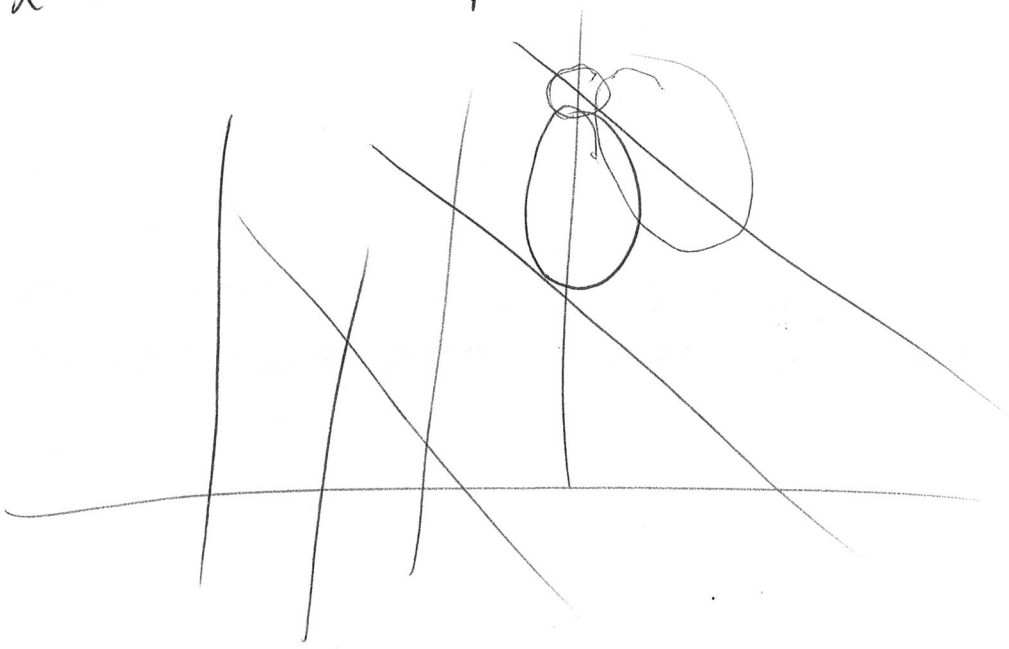
Score based Likelihood Ratio / Bayes factor

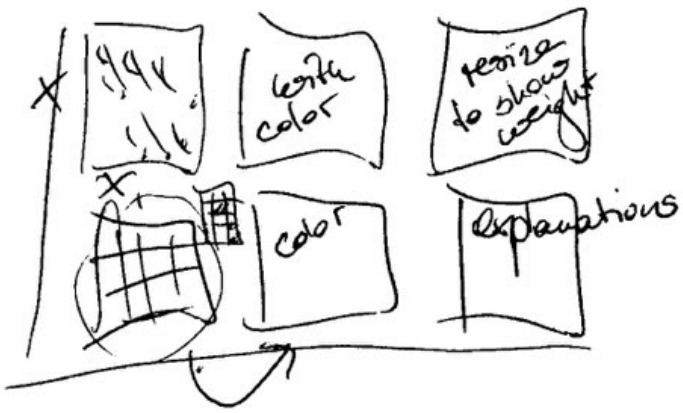
global summary



$$z = ax_1 + bx_2$$

local summaries





feature selection (not shown)

Create permutations

obtain RF predictions

weight (gower)

bin

weight (not gower)

fit ridge regression

~~get expla~~

Select features

re-fit ridge

get explanations