Interpretable Machine Learning with R

Dean Allsopp

SatRday Newcastle

April 6, 2019

Why “Interpretable” Machine Learning?

In the beginning…

Linear Models

Let there be light…

- Coefficients
- Residual plots
- QQ plots
- Leverage plots

Kaggle Era DS…

It’s all about the error score

Darkness falls…

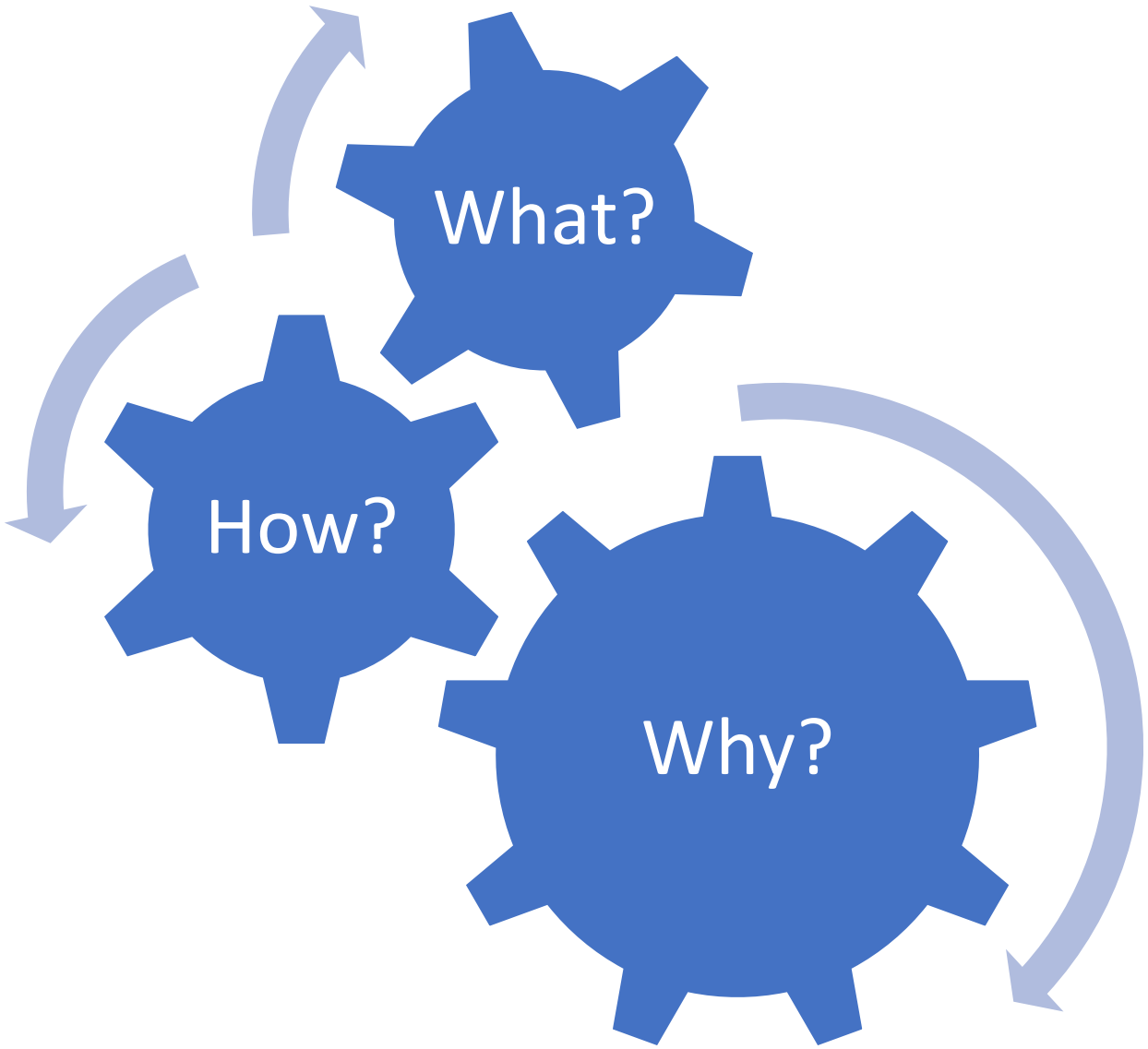
- Complex Features
- Complex Models
- Hyperparameter Search

But what is it actually doing to predict this result?

Just hypothetically important?

What constitutes the best medical treatment for recovery?

See Rich Caruana

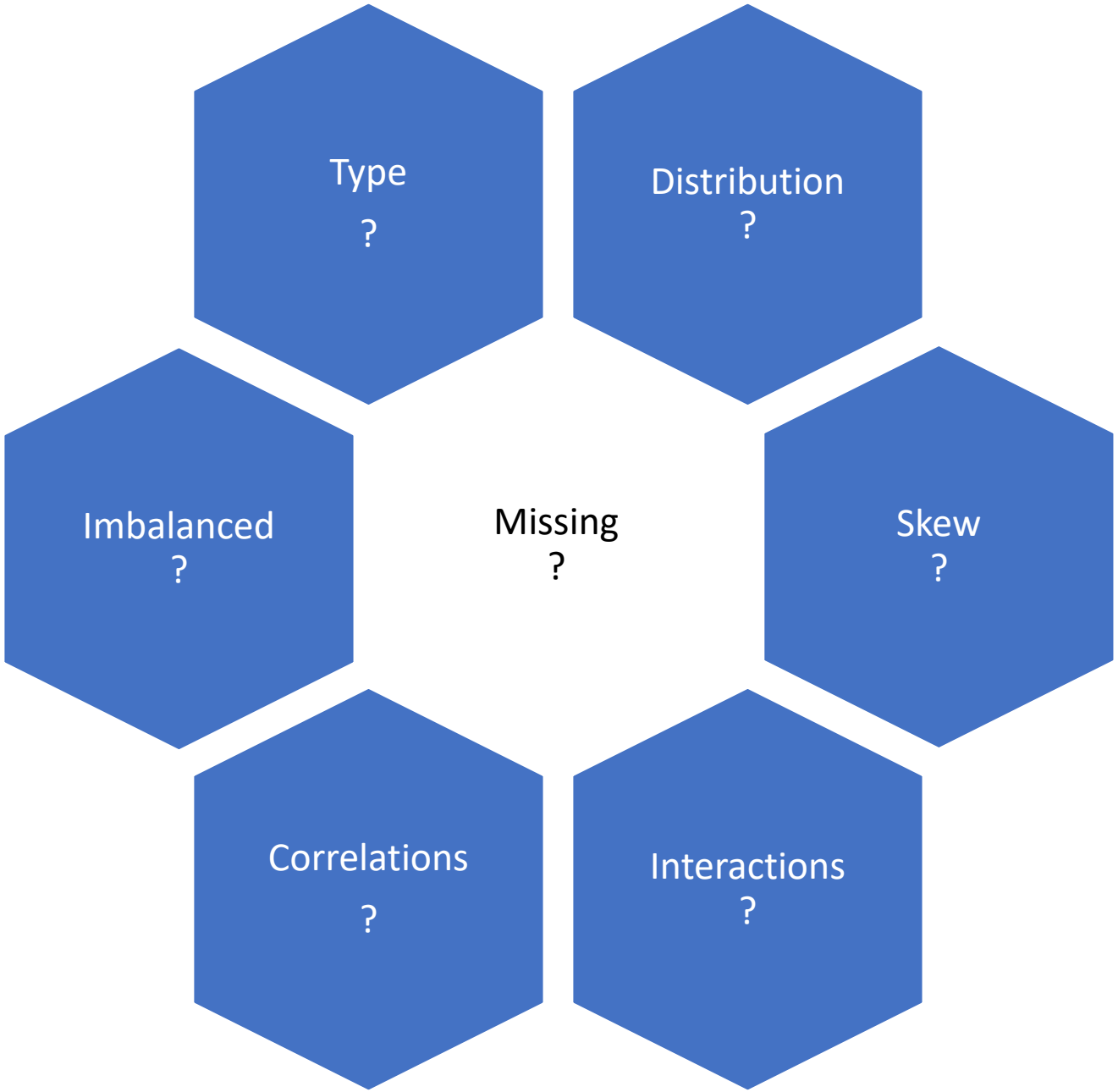

What?
How?
Why?

What? – Data (Features/Instances)

How? – does the Model work?

Why? – do we get this Prediction?

How well do you understand the data?

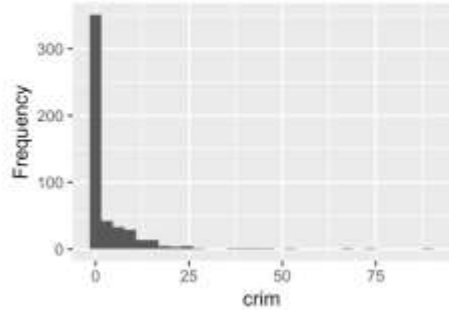

Type
?
Distribution
?
Imbalanced
?
Missing
?
Skew
?
Correlations
?
Interactions
?

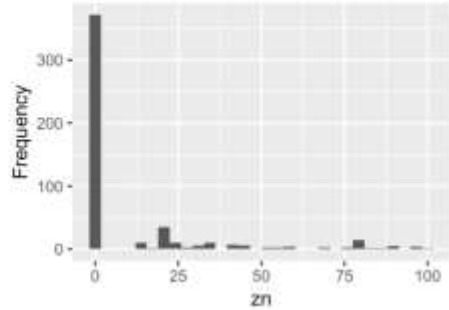
EDA

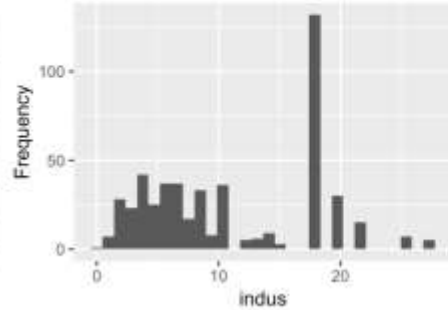
Data Explorer

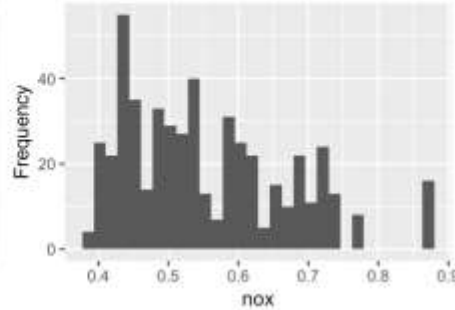
Univariate Distribution

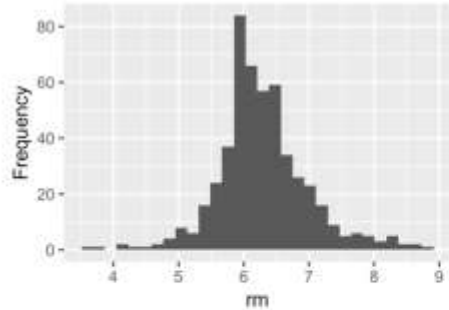
Histogram

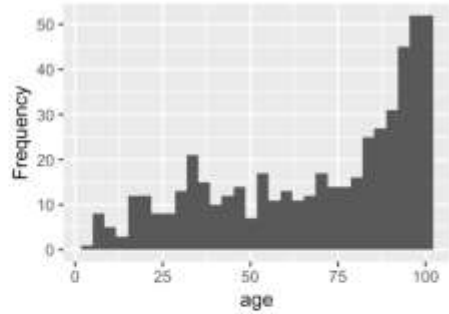


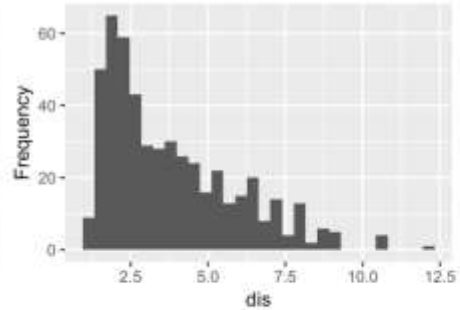


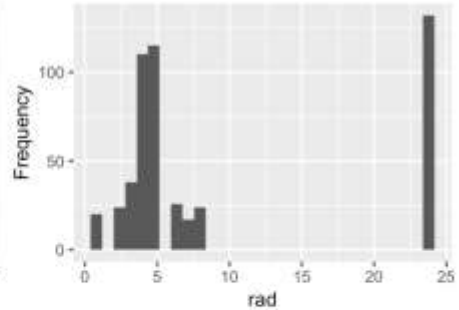


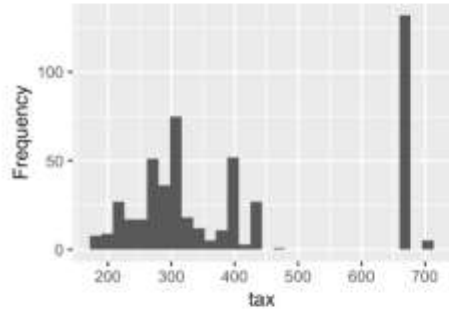


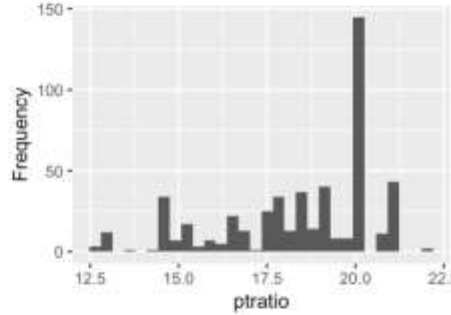


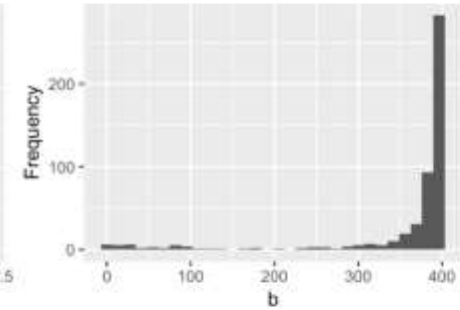


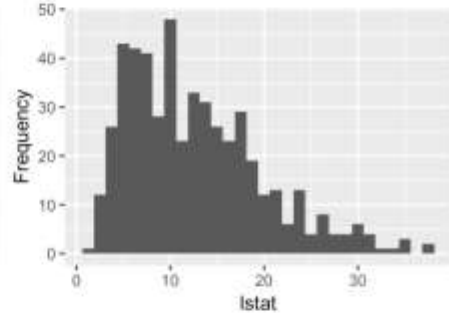


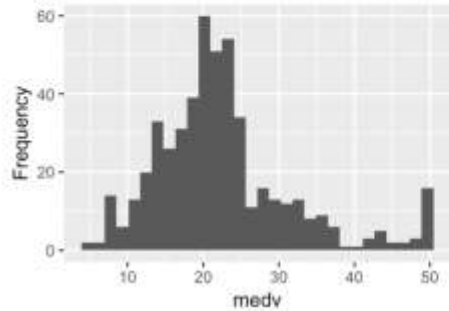




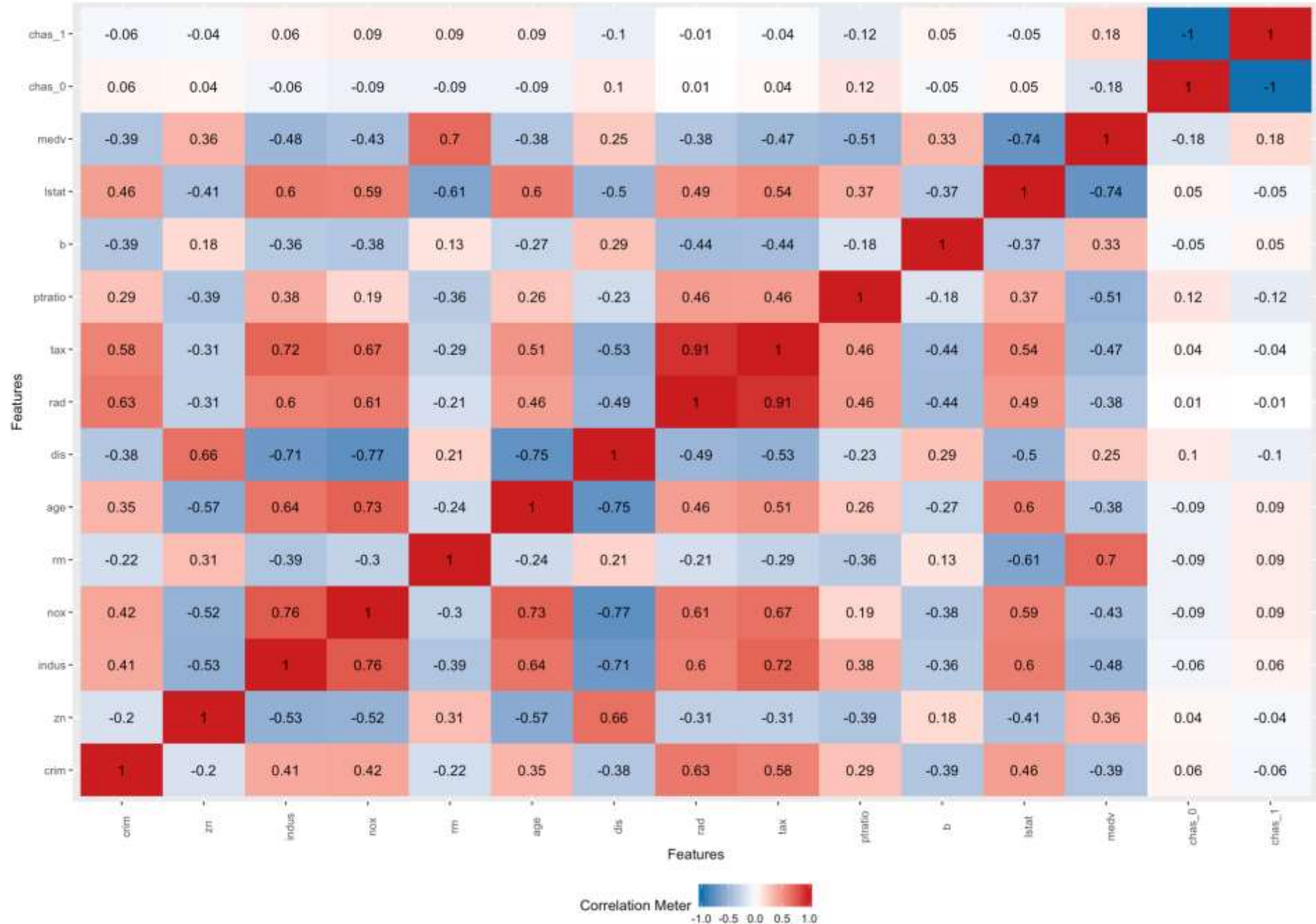




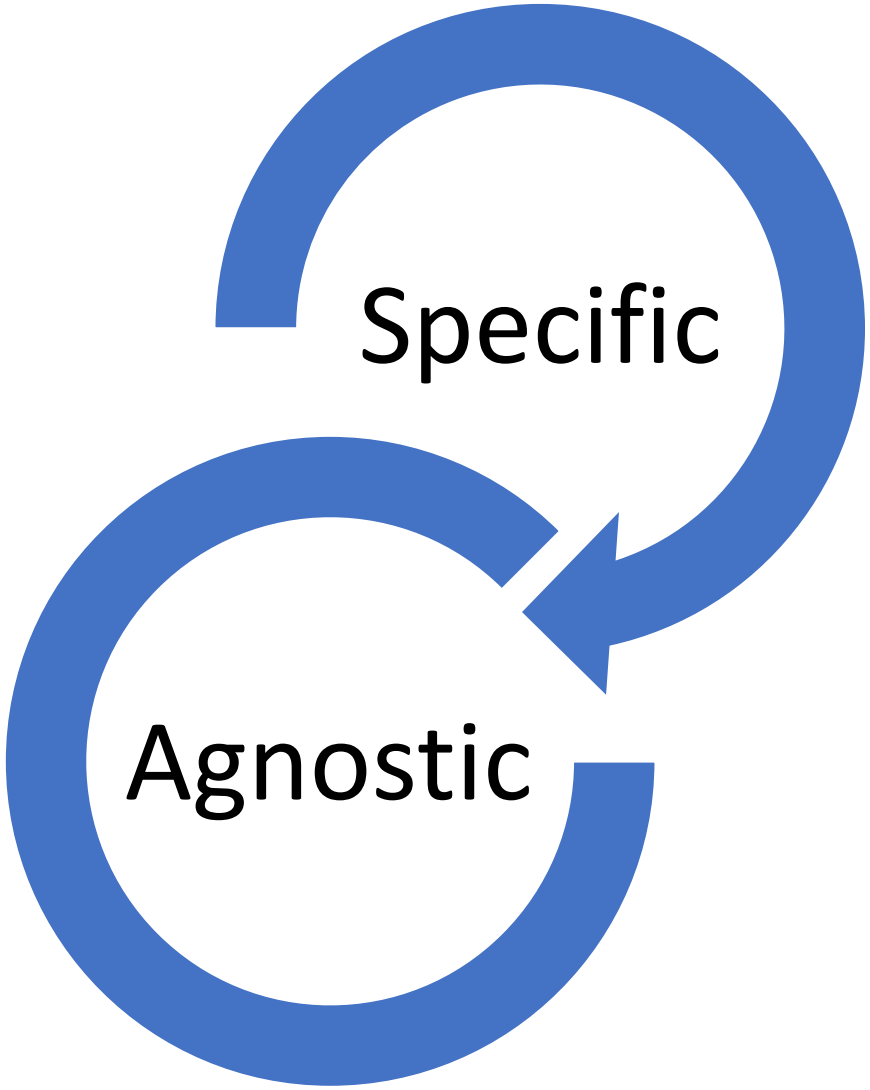




Correlation Analysis



Kind of Model Interpretation?

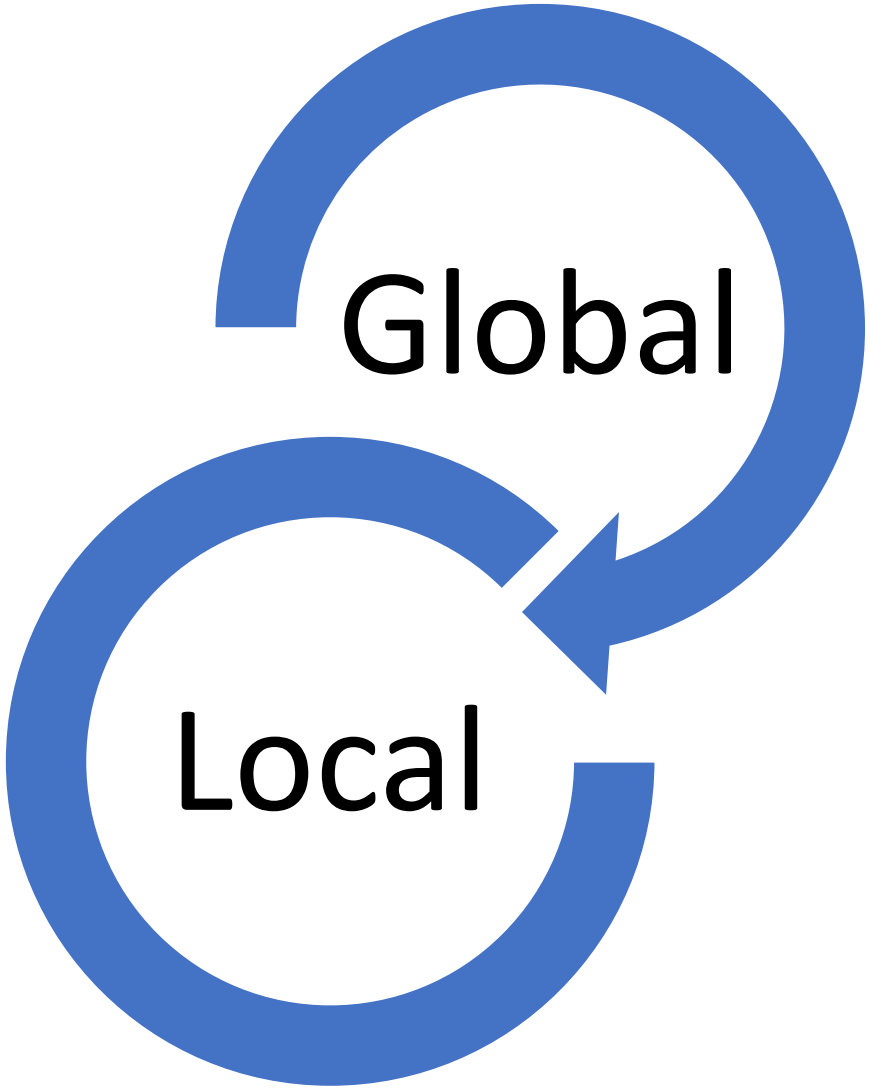

Specific
Agnostic

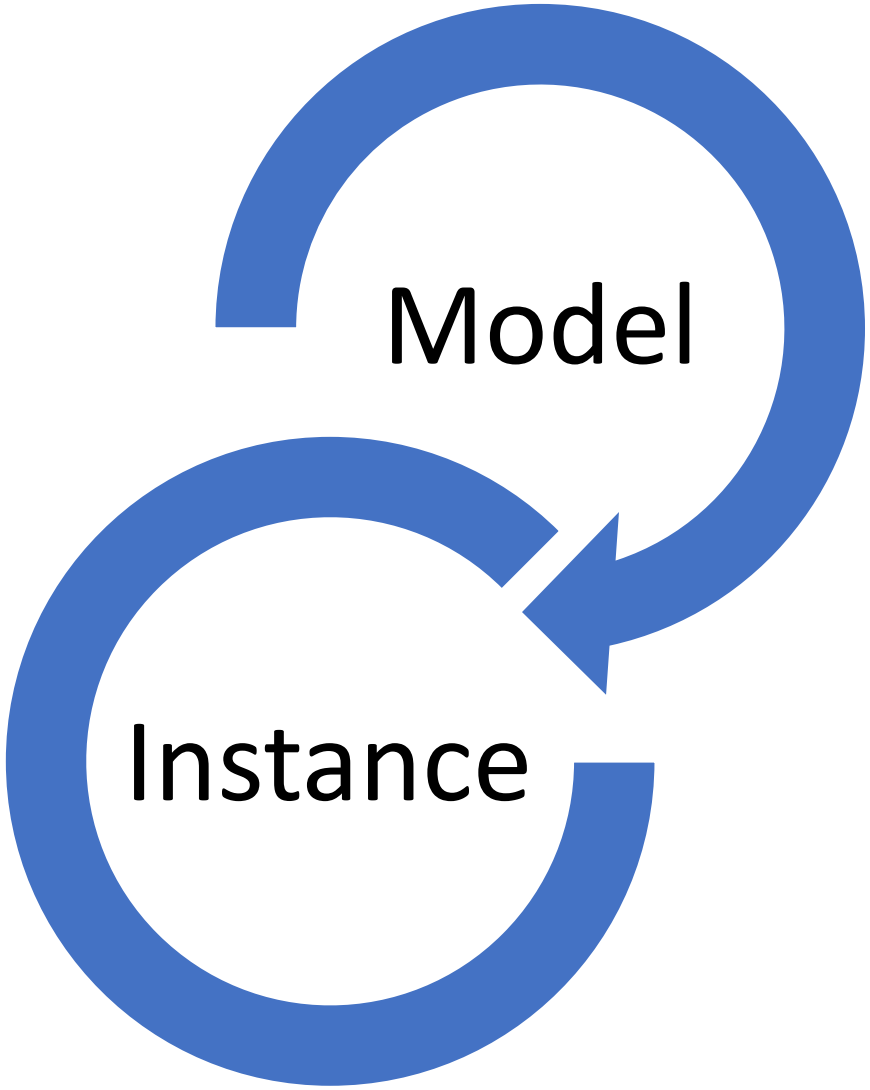
Exploit Internal Logic

Versus

System Identification Approach

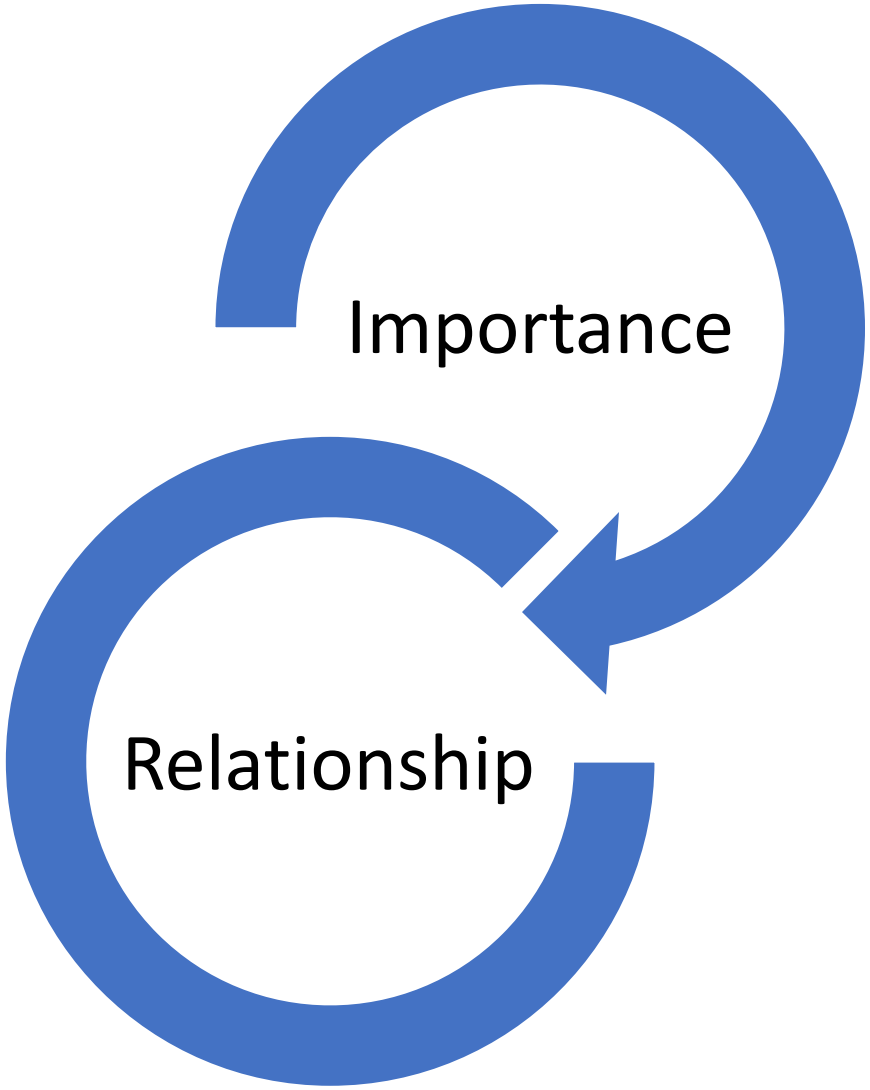
Scope of Model Interpretation?


Global
Local


Model
Instance

#1: Global

What do you want to understand about the features and the model?


Importance
Relationship

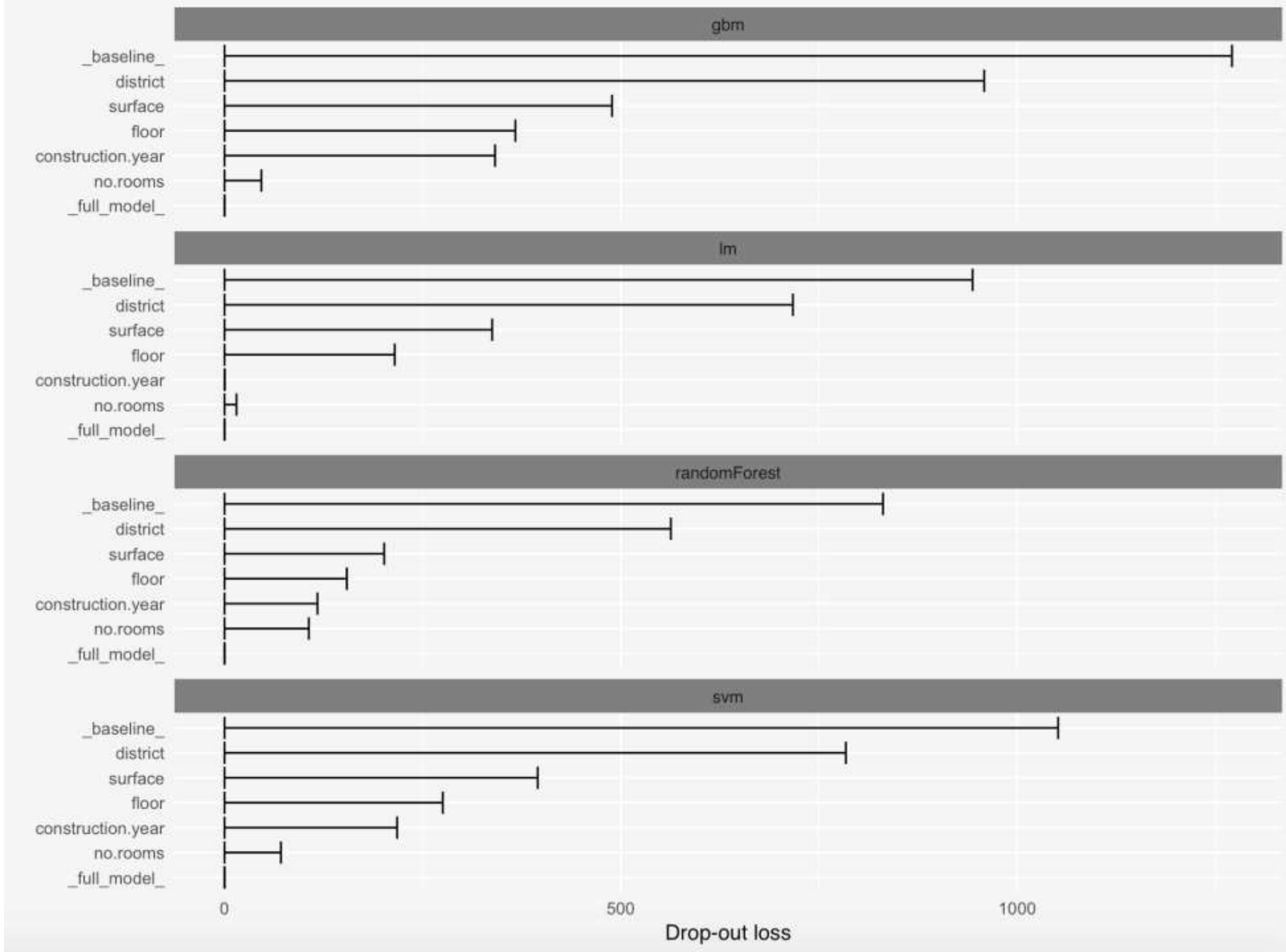
Variable Importance Plot

DALEX/VIP (agnostic)

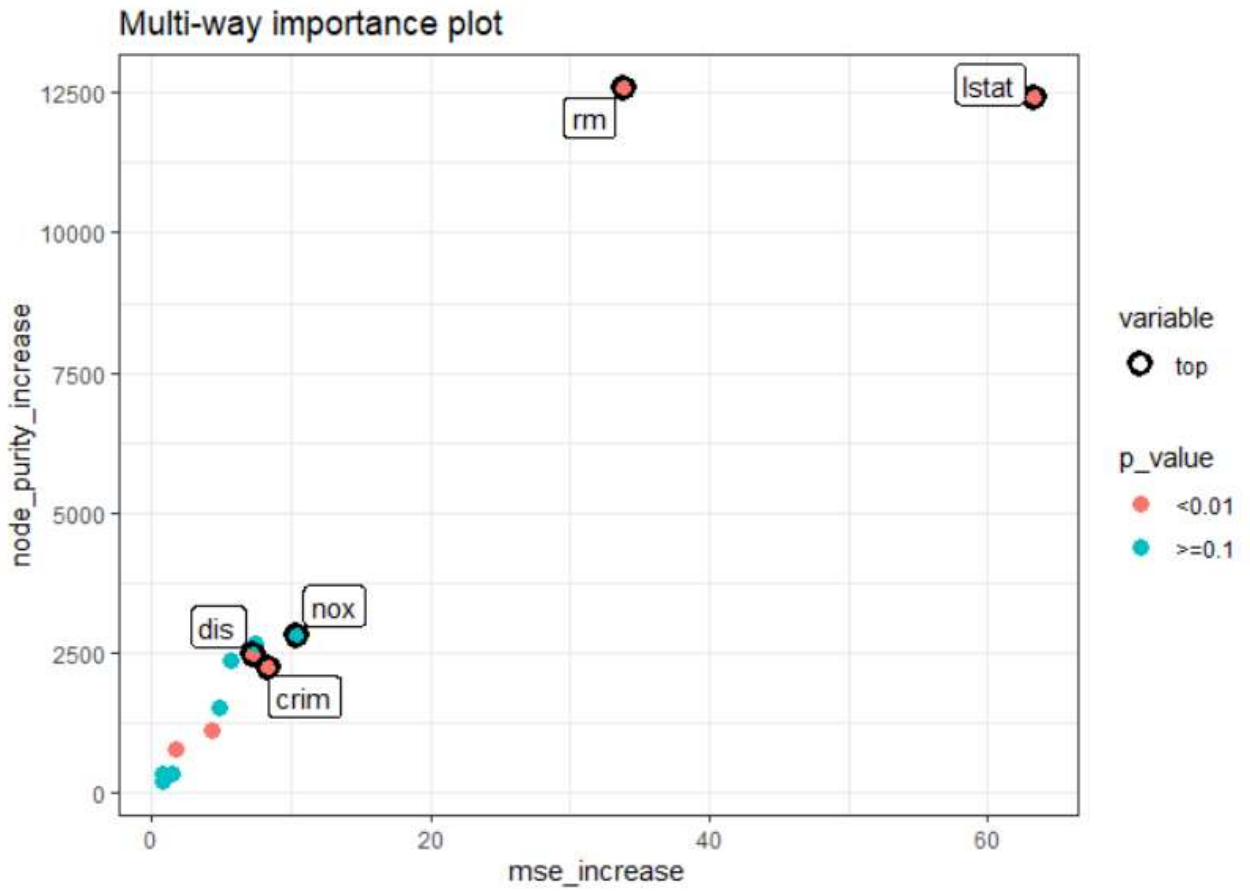
randomForestExplainer

Why highlight VIP?

Permutation

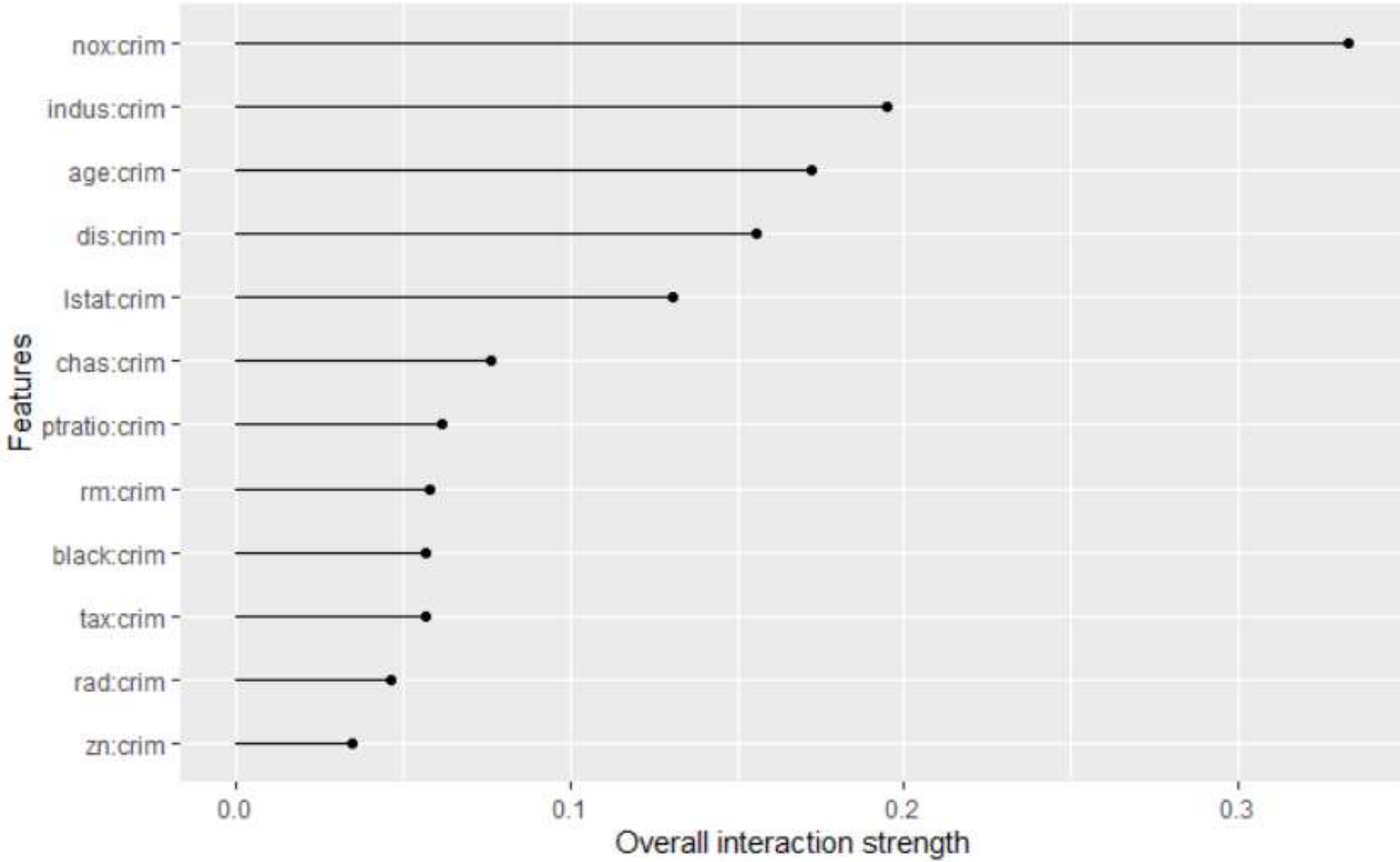

gbm
baseline
district
surface
floor
construction.year
no.rooms
_full_model_
lm
baseline
district
surface
floor
construction.year
no.rooms
_full_model_
randomForest
baseline
district
surface
floor
construction.year
no.rooms
_full_model_
svm
baseline
district
surface
floor
construction.year
no.rooms
_full_model_
0
500
1000
Drop-out loss

randomForestExplainer



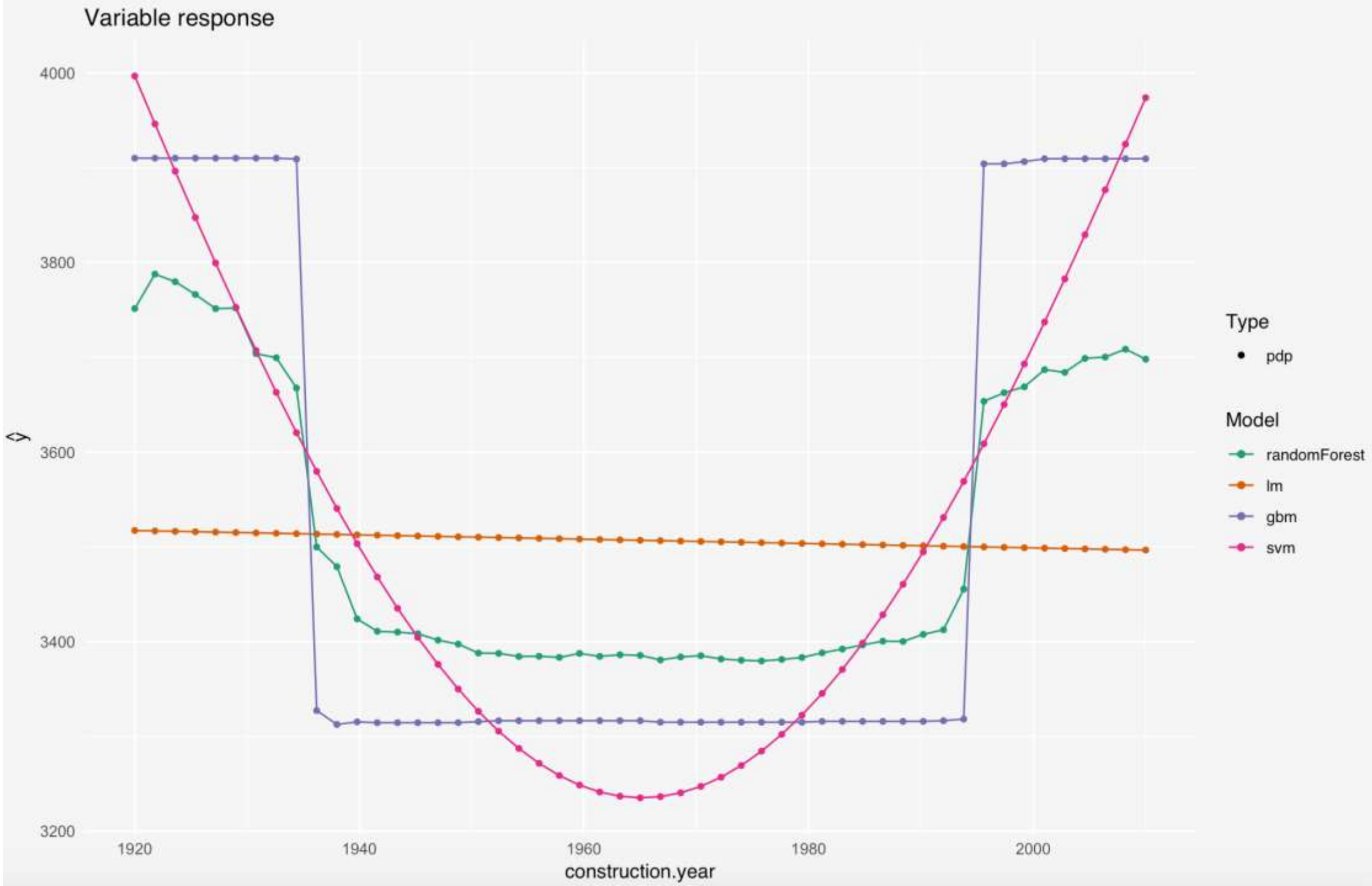
What about Interactions?

IML


nox:crim
indus:crim
age:crim
dis:crim
lstat:crim
chas:crim
ptratio:crim
rm:crim
black:crim
tax:crim
rad:crim
zn:crim
Features
0.0
0.1
0.2
0.3
Overall interaction strength

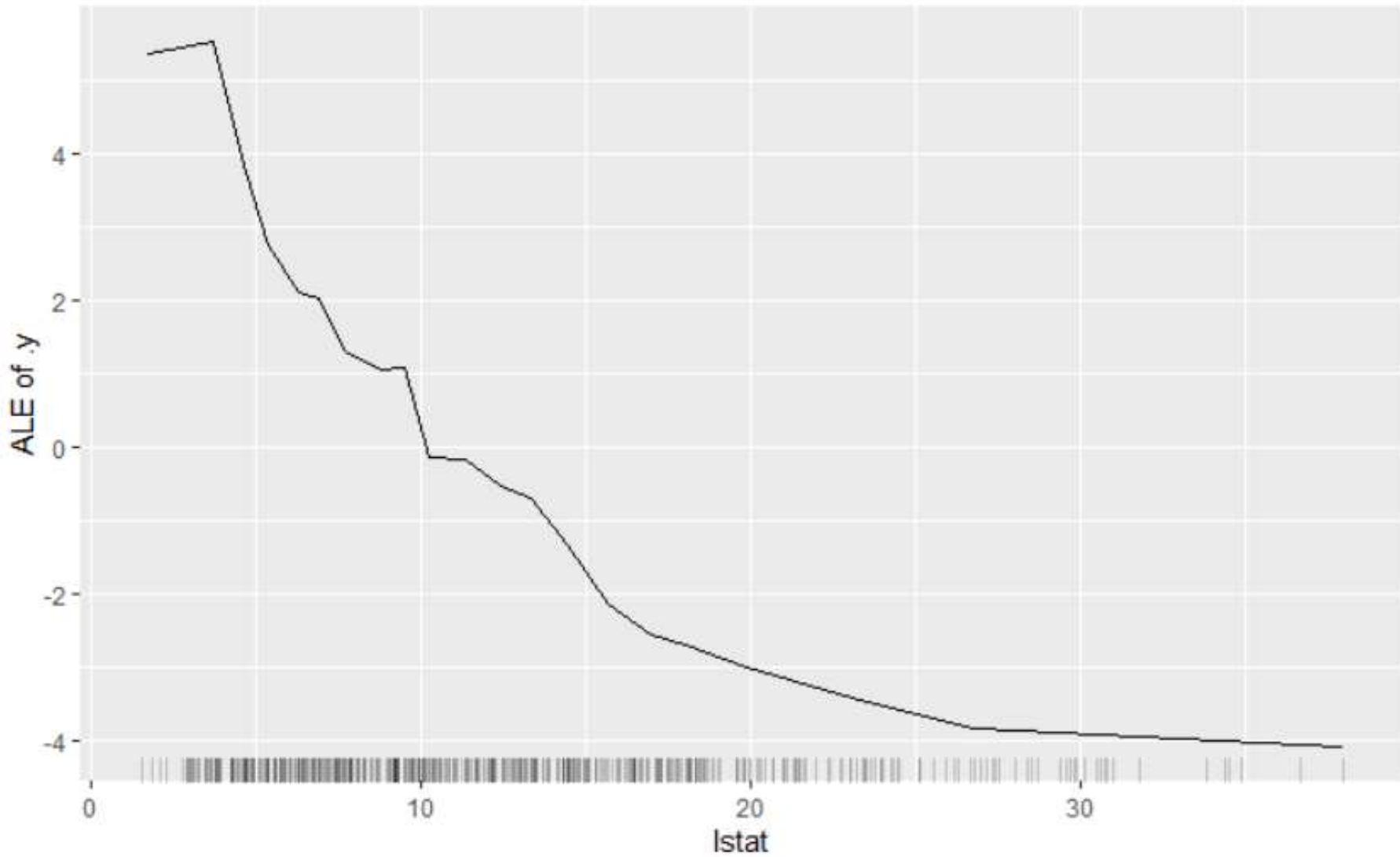
Partial Dependence Plot

DALEX/IML/plotmo


Variable response
4000
3800
3600
3400
3200
<y>
1920
1940
1960
1980
2000
construction.year
Type
pdp
Model
randomForest
lm
gbm
svm

Accumulated Local Effects Plot

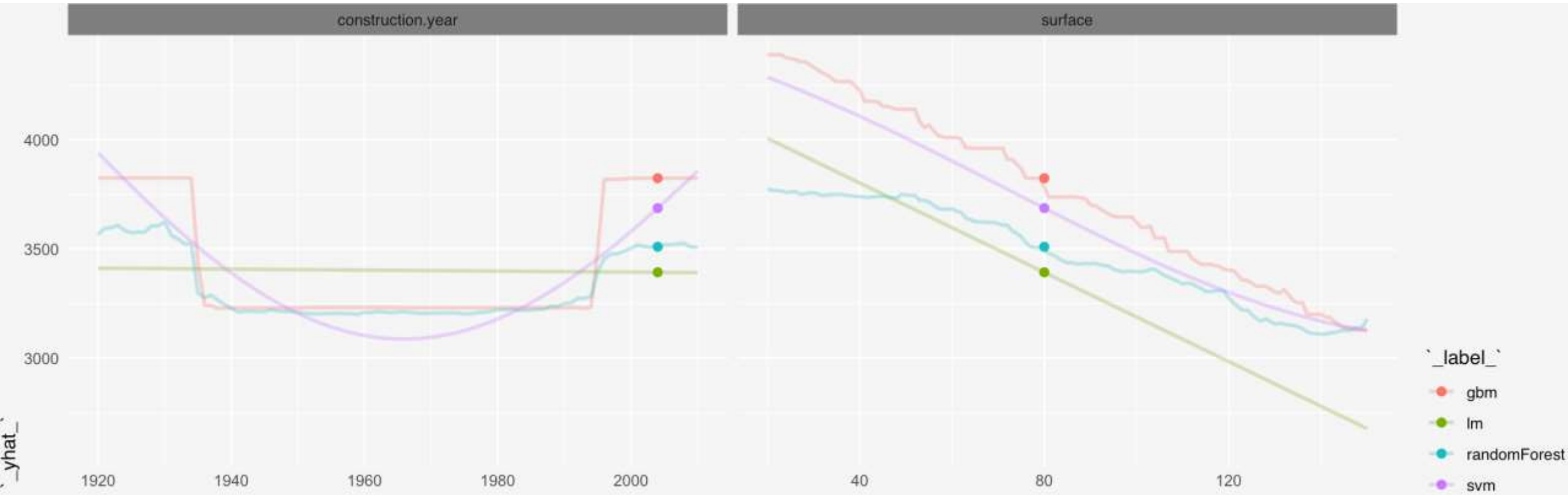
DALEX/IML


ALE of .y
4
2
0
-2
-4
0
10
20
30
lstat

What If?

Ceteris Paribus Plot

DALEX


construction.year
surface
4000
3500
3000
`_yhat_`
1920
1940
1960
1980
2000
40
80
120
`_label_`
gbm
lm
randomForest
svm

#2: Local

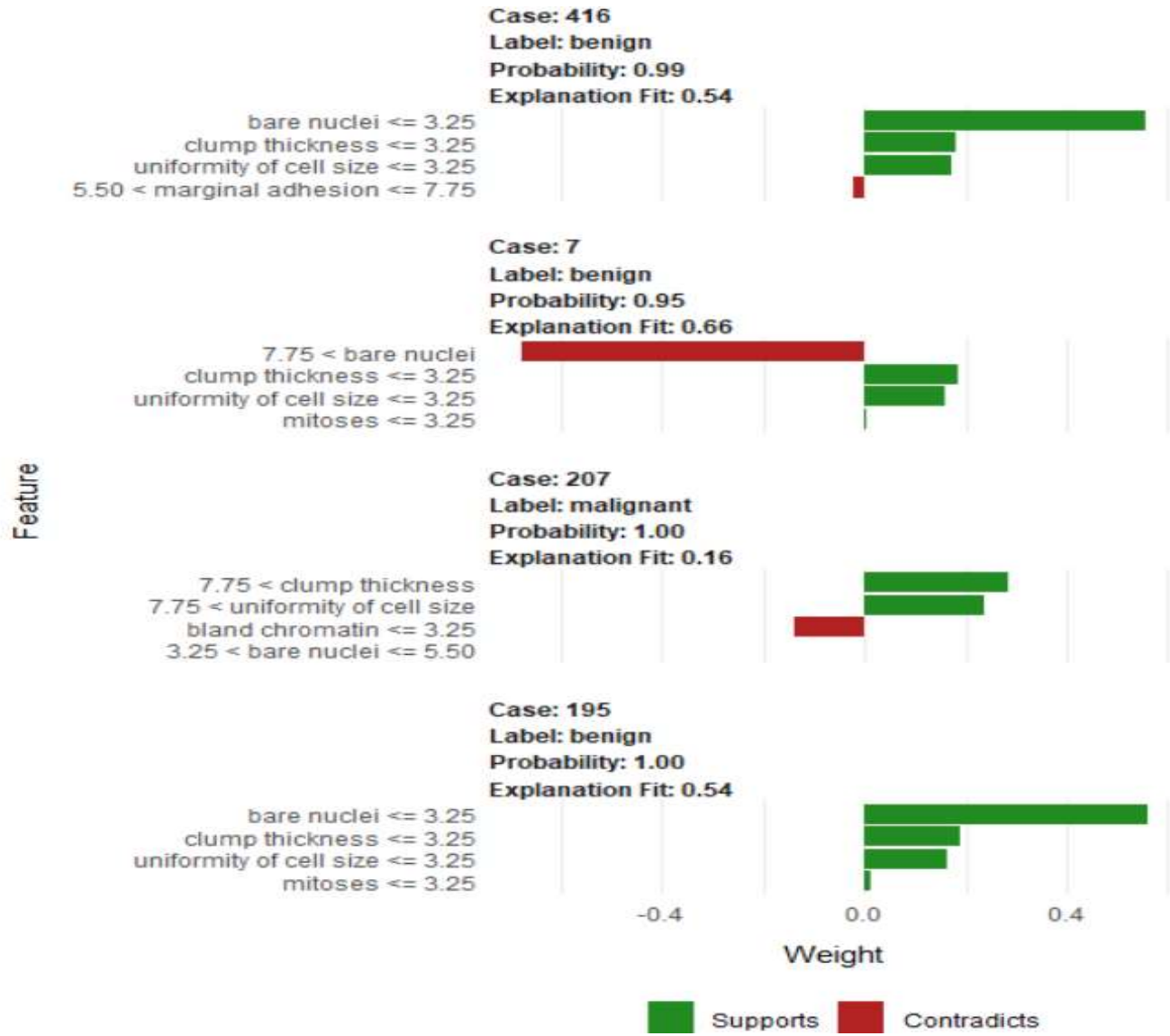
Local Interpretable Model-Agnostic Explanations

lime

Generate data points based on training data

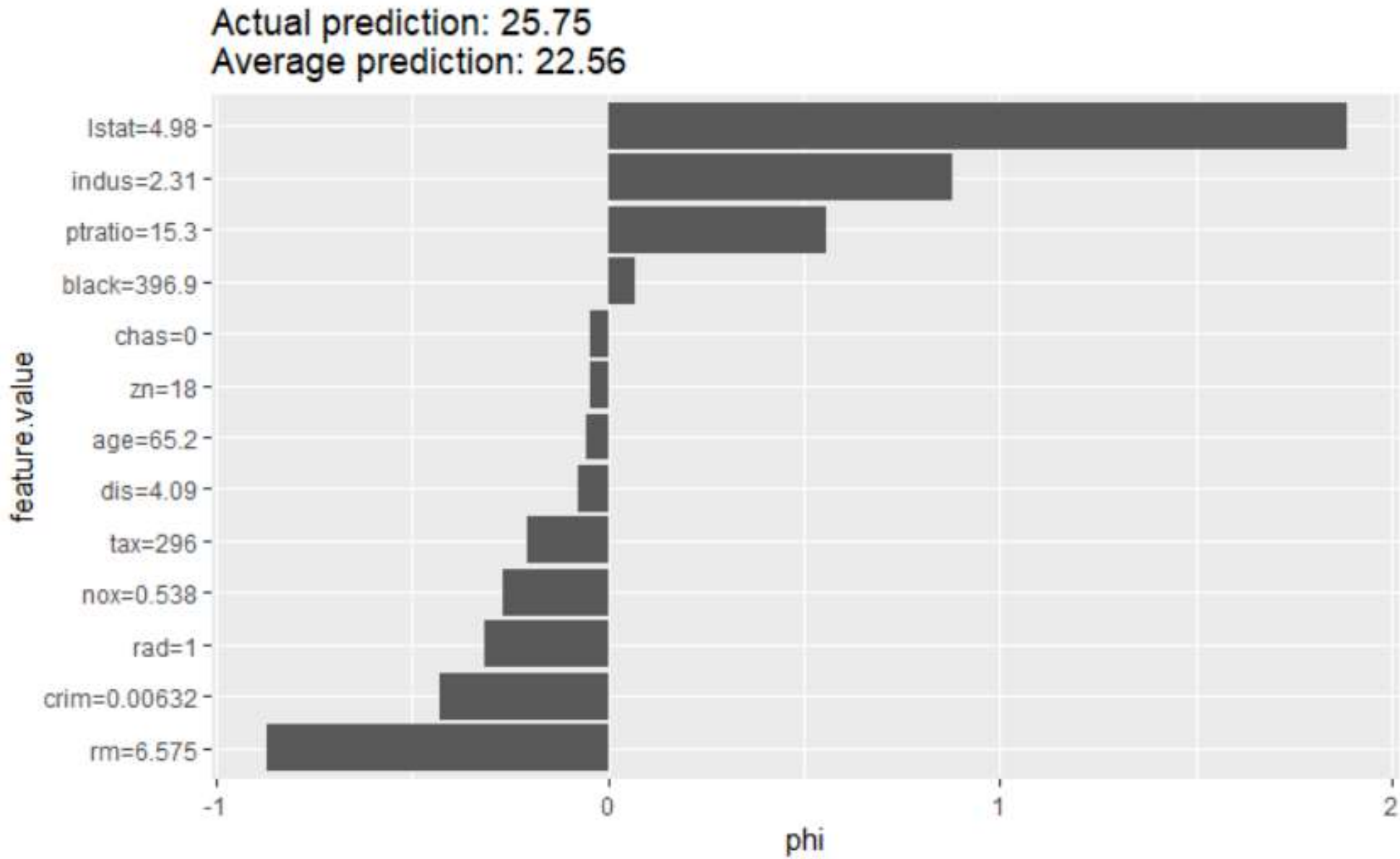
Compute complex model predictions from the generated data to find the 'most useful features'

Fit a local linear model for the 'most useful features' and use the feature coefficients as reason codes


Case: 416
Label: benign
Probability: 0.99
Explanation Fit: 0.54
bare nuclei <= 3.25
clump thickness <= 3.25
uniformity of cell size <= 3.25
5.50 < marginal adhesion <= 7.75
Case: 7
Label: benign
Probability: 0.95
Explanation Fit: 0.66
7.75 < bare nuclei
clump thickness <= 3.25
uniformity of cell size <= 3.25
mitoses <= 3.25
Case: 207
Label: malignant
Probability: 1.00
Explanation Fit: 0.16
7.75 < clump thickness
7.75 < uniformity of cell size
bland chromatin <= 3.25
3.25 < bare nuclei <= 5.50
Case: 195
Label: benign
Probability: 1.00
Explanation Fit: 0.54
bare nuclei <= 3.25
clump thickness <= 3.25
uniformity of cell size <= 3.25
mitoses <= 3.25
Feature
-0.4
0.0
0.4
Weight
Supports
Contradicts

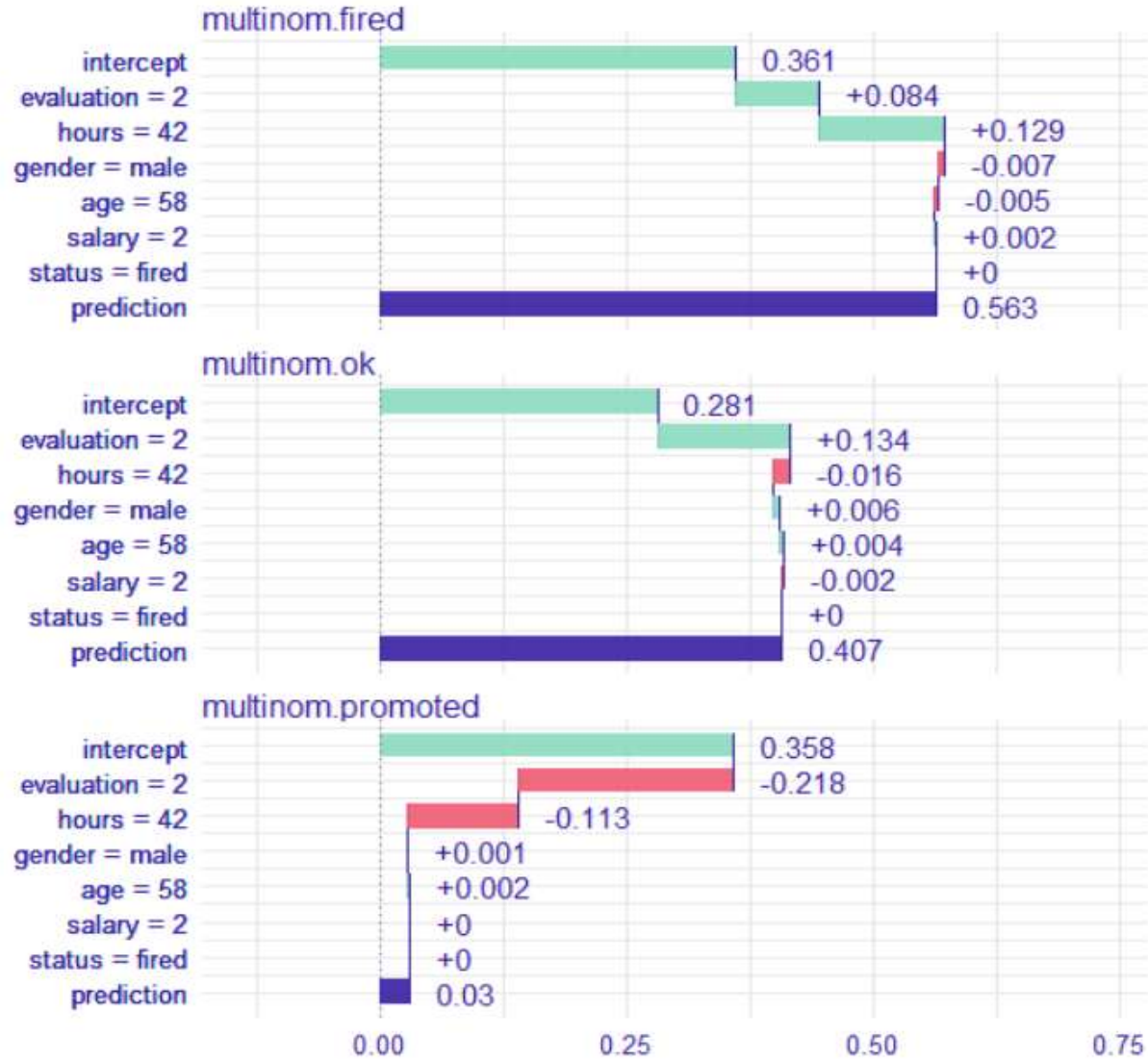
Shapley Values (Coalition Attribution)

IML/ShapleyR/Shapper/ ExplainPrediction


Actual prediction: 25.75
Average prediction: 22.56
lstat=4.98
indus=2.31
ptratio=15.3
black=396.9
chas=0
zn=18
age=65.2
dis=4.09
tax=296
nox=0.538
rad=1
crim=0.00632
rm=6.575
feature.value
-1
0
1
2
phi

BreakDown
(Feature Contributions)

iBreakDown*


multinom.fired
intercept
evaluation = 2
hours = 42
gender = male
age = 58
salary = 2
status = fired
prediction
0.361
+0.084
+0.129
-0.007
-0.005
+0.002
+0
0.563
multinom.ok
intercept
evaluation = 2
hours = 42
gender = male
age = 58
salary = 2
status = fired
prediction
0.281
+0.134
-0.016
+0.006
+0.004
-0.002
+0
0.407
multinom.promoted
intercept
evaluation = 2
hours = 42
gender = male
age = 58
salary = 2
status = fired
prediction
0.358
-0.218
-0.113
+0.001
+0.002
+0
+0
0.03
0.00
0.25
0.50
0.75

Gotchas?

Stability?

Fidelity?

Scope?

Foundations

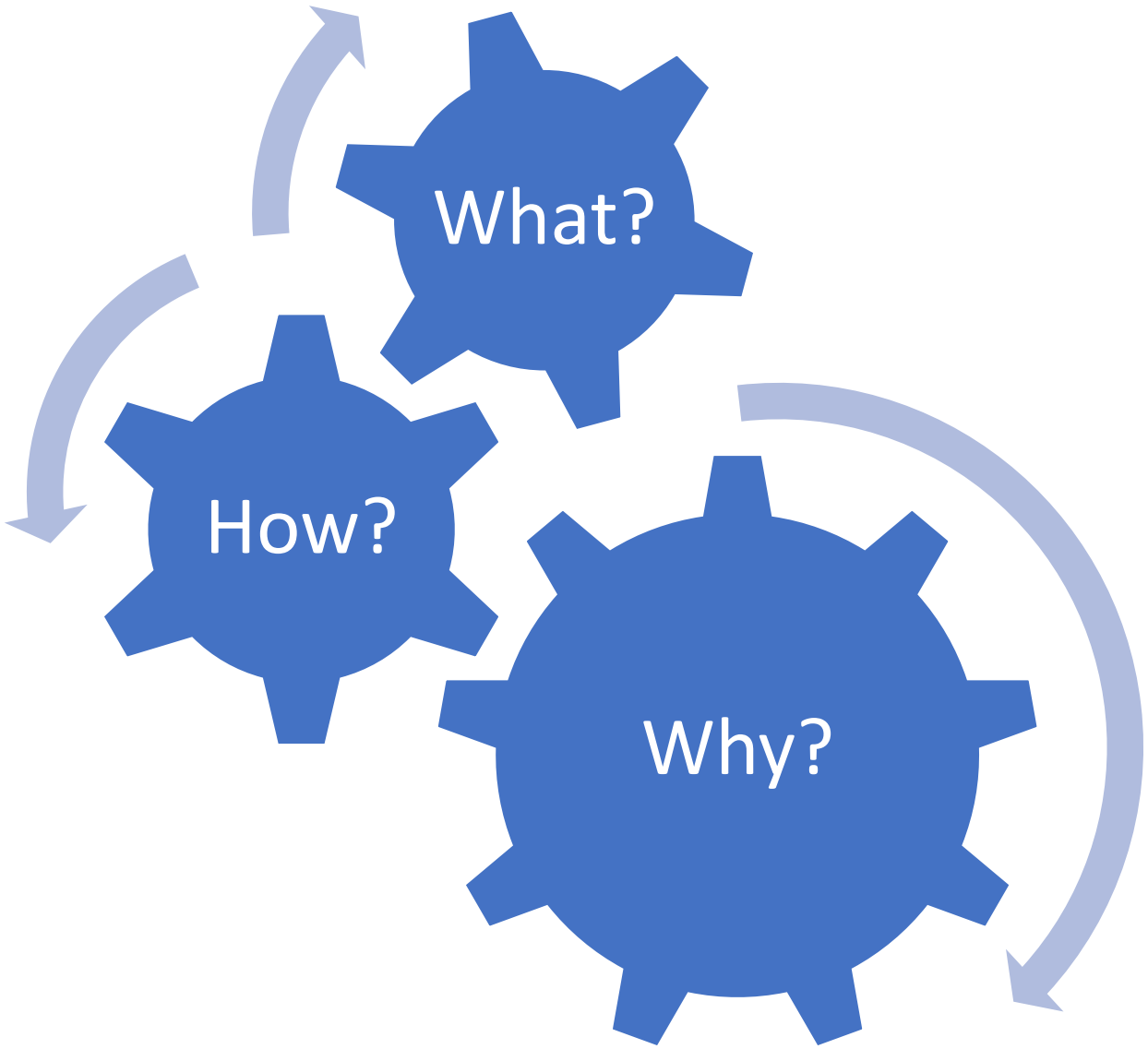
Interpretable Data
Performant Model

“Variable Importances”…
Linearity Assumptions…
Collinearity
Computation Time

Conclusion

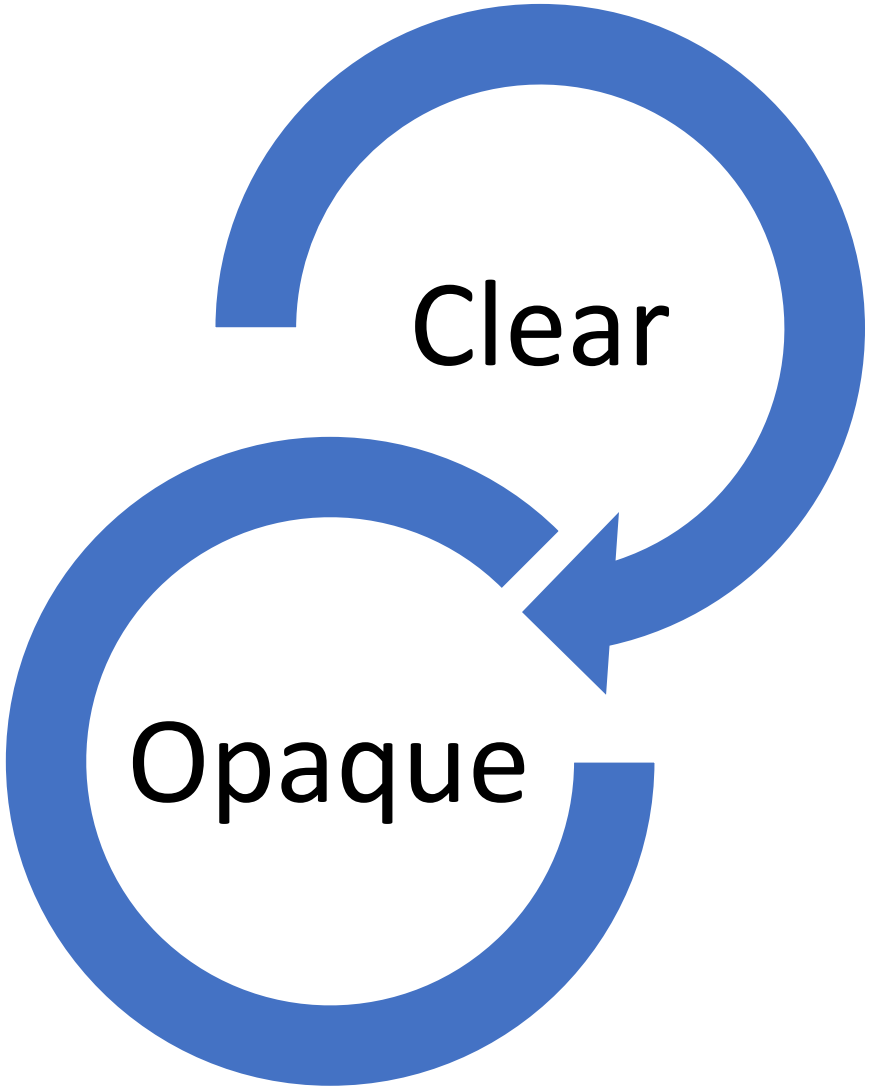
Compare different ML Models

Compare different Interpretation Approaches


What?
How?
Why?

Consider Local and Global

Combine Multiple Perspectives

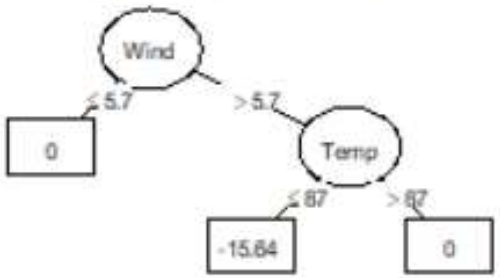

Clear
Opaque

More transparent models?

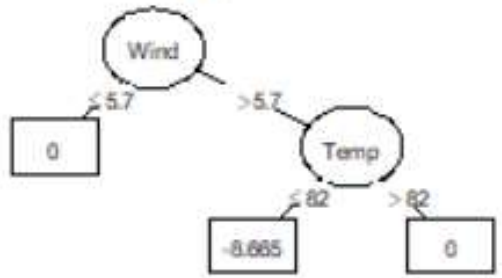
Rule Ensembles

pre

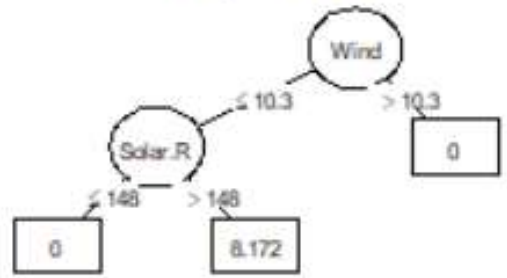
rule191: Importance = 6.263



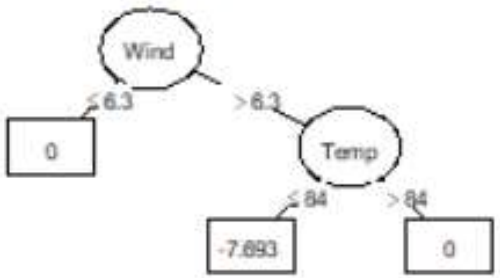
rule173: Importance = 4.074



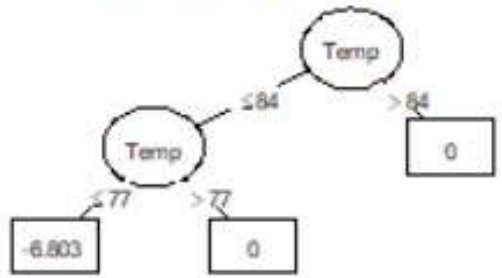
rule204: Importance = 4.056



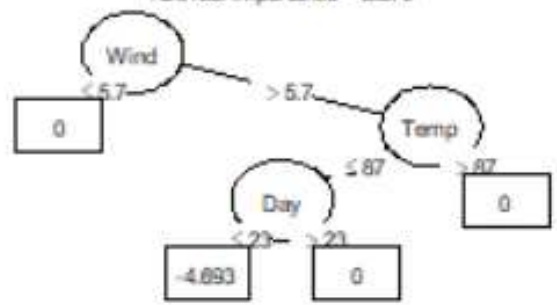
rule42: Importance = 3.591



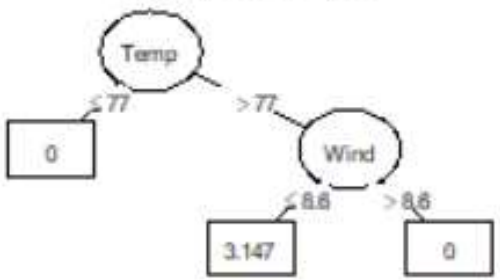
rule10: Importance = 3.4



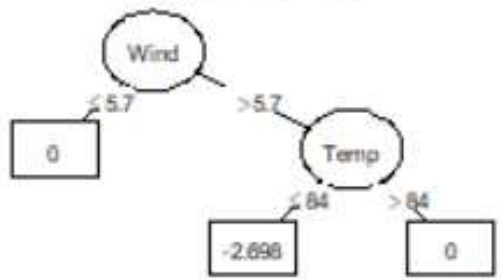
rule192: Importance = 2.275



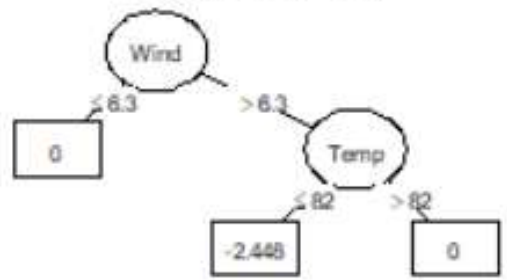
rule93: Importance = 1.457



rule51: Importance = 1.228



rule25: Importance = 1.167



Recommendations:

Data - [Data Explorer]
Local - Shapley Values [IML]
Global - Variable Importance [DALEX]
Plots - ALE, Ceteris Paribus [DALEX]
Plots – ALE, PDP, ICE [IML]

Thanks for Listening!

dean_allsopp@hotmail.com