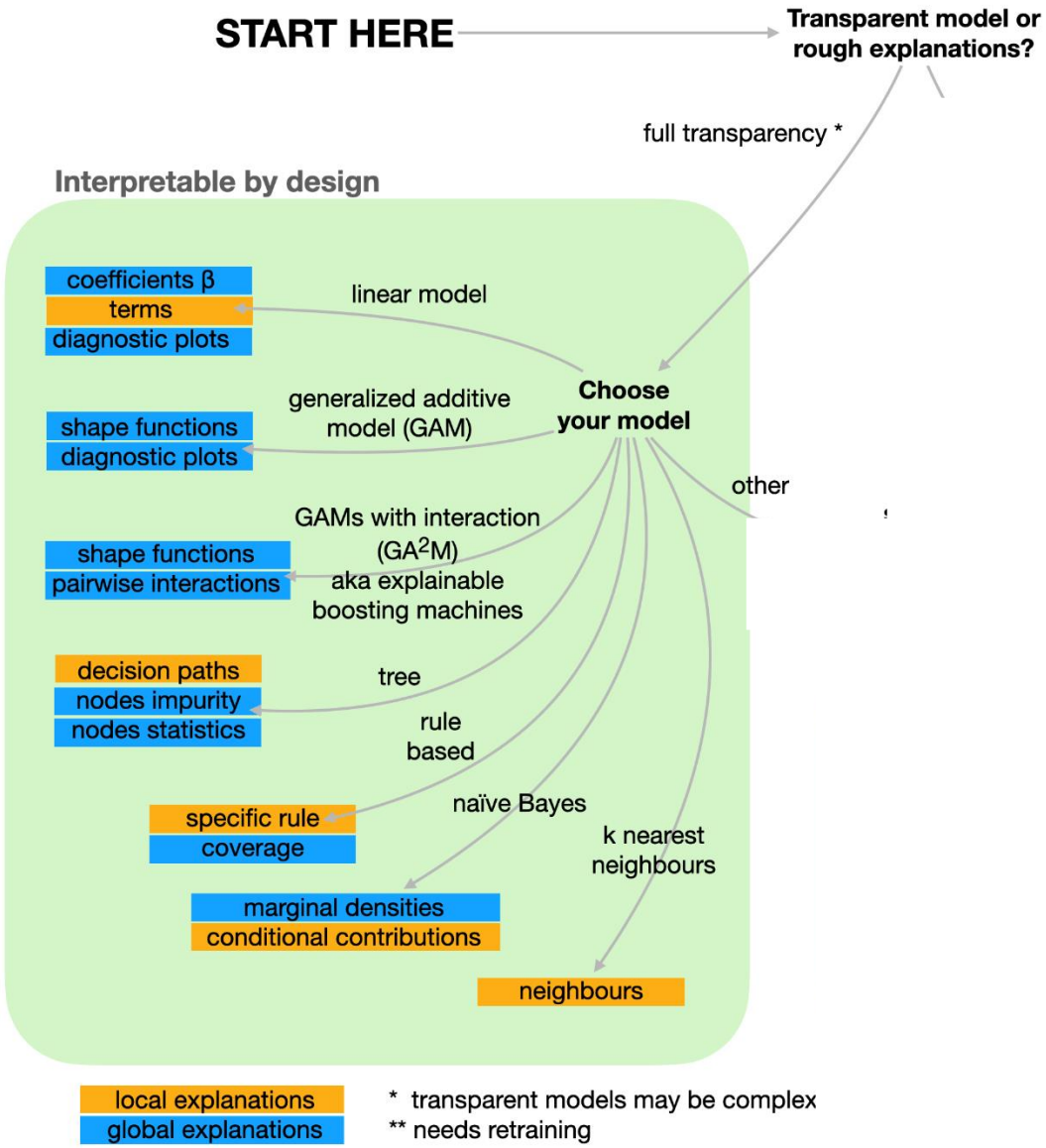
Landscape of R packages for eXplainable Artificial Intelligence

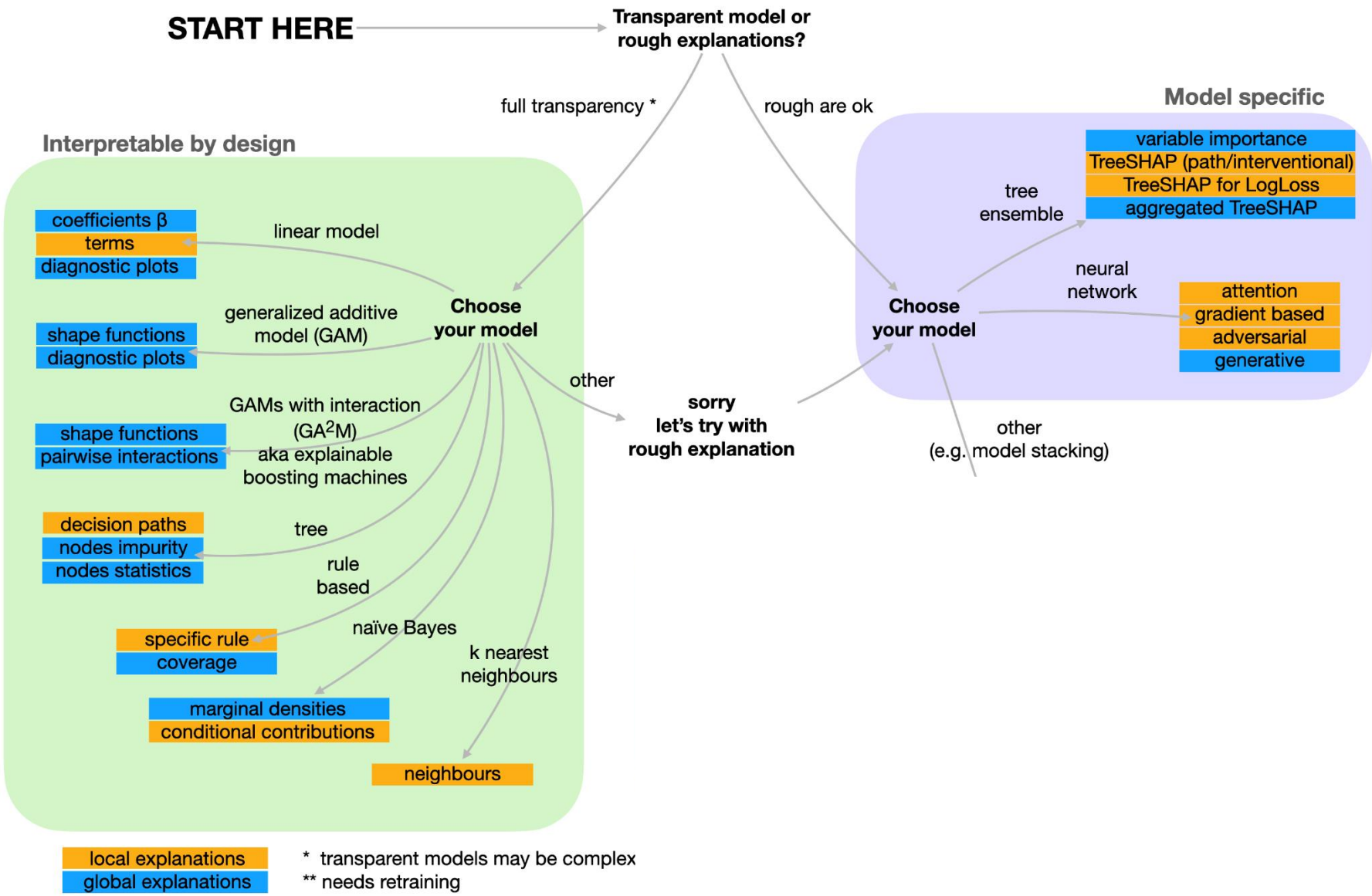
Szymon Maksymiuk and Alicja Gosiewska

12.10.2020

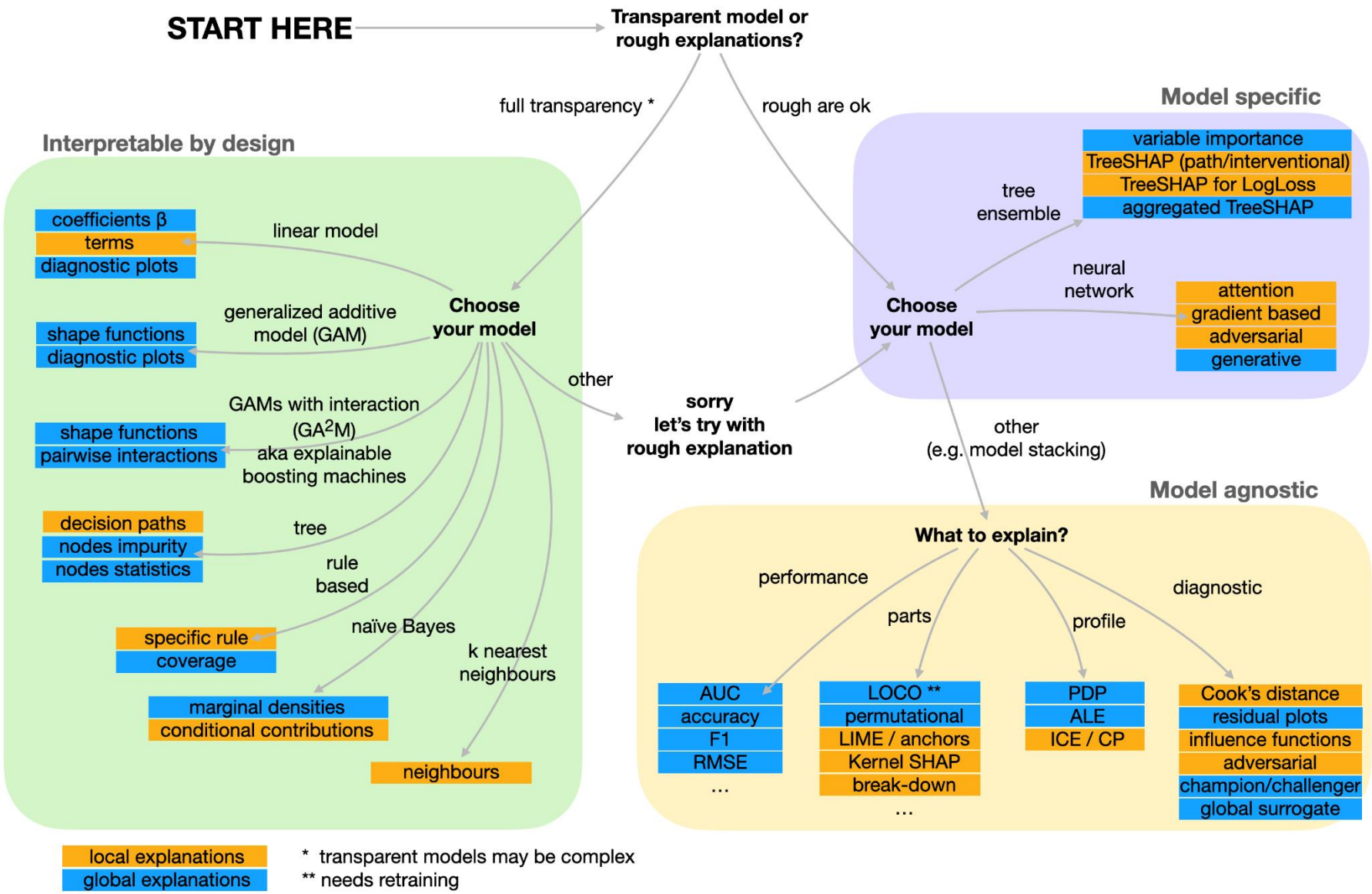
Taxonomy in terms of models



Taxonomy in terms of models



Taxonomy in terms of models



Taxonomy in terms of explanation object

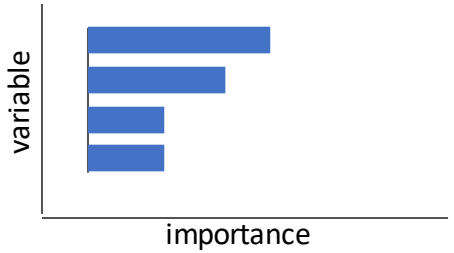
Global Explanations

Local Explanations

Taxonomy in terms of explanation object

Global Explanations

Model Parts



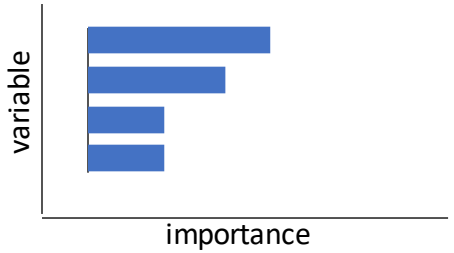
Permutational feature importance
Leave One Covariate Out (LOCO)
Surrogate models
Aggreagated SHapley Additive exPlanations

Local Explanations

Taxonomy in terms of explanation object

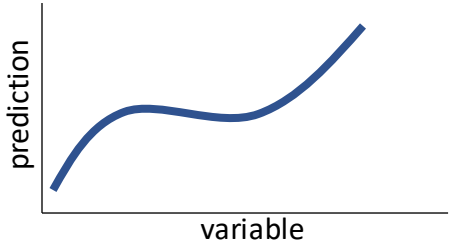
Global Explanations

Model Parts



Permutational feature importance
Leave One Covariate Out (LOCO)
Surrogate models
Aggreagated SHapley Additive exPlanations

Model Profile



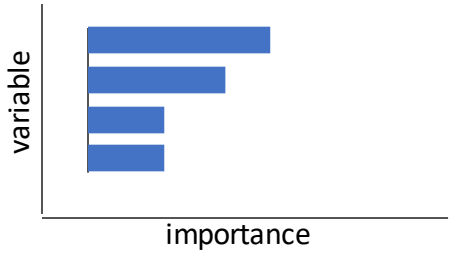
Partial Dependance Profiles (PDP)
Accumulated Local Effects (ALE)

Local Explanations

Taxonomy in terms of explanation object

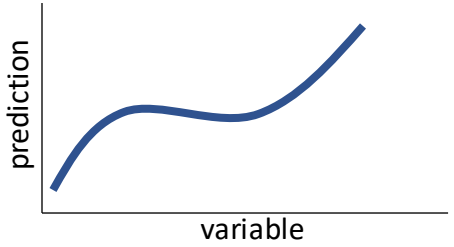
Global Explanations

Model Parts



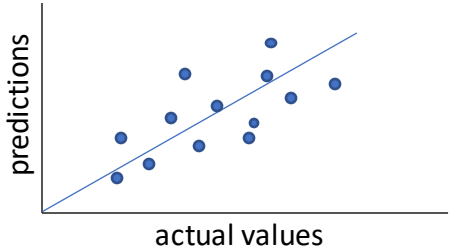
Permutational feature importance
Leave One Covariate Out (LOCO)
Surrogate models
Aggreagated SHapley Additive exPlanations

Model Profile



Partial Dependance Profiles (PDP)
Accumulated Local Effects (ALE)

Model Diagnostics



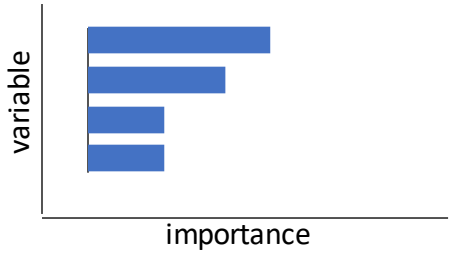
Residual plots
Variable vs. prediction plots
Demographic parity

Local Explanations

Taxonomy in terms of explanation object

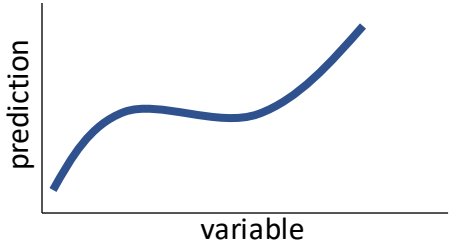
Global Explanations

Model Parts



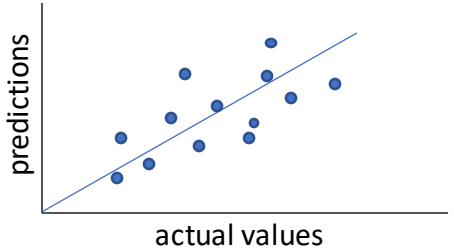
Permutational feature importance
Leave One Covariate Out (LOCO)
Surrogate models
Aggreagated SHapley Additive exPlanations

Model Profile



Partial Dependance Profiles (PDP)
Accumulated Local Effects (ALE)

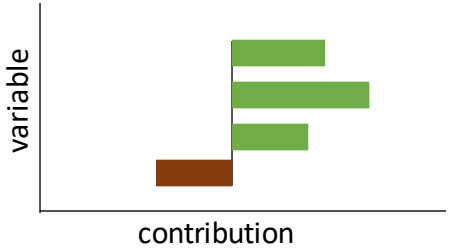
Model Diagnostics



Residual plots
Variable vs. prediction plots
Demographic parity

Local Explanations

Predict Parts

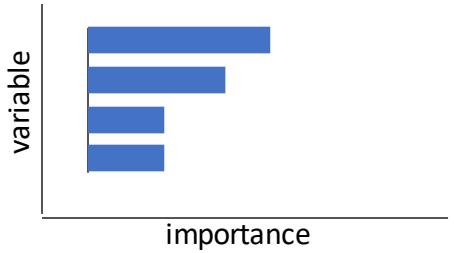


Break Down (BD)
SHapley Additive exPlanations (SHAP)
Local Interpretable Model agnostic Explanations (LIME)

Taxonomy in terms of explanation object

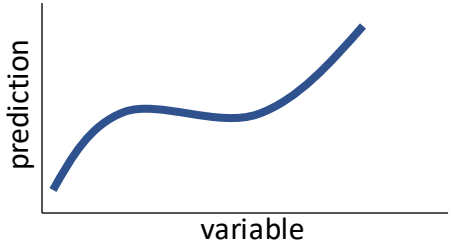
Global Explanations

Model Parts



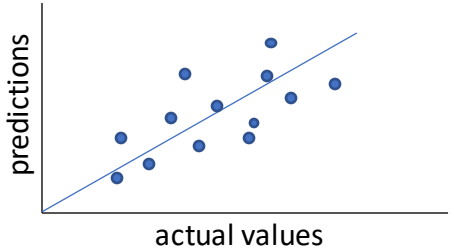
Permutational feature importance
Leave One Covariate Out (LOCO)
Surrogate models
Aggreagated SHapley Additive exPlanations

Model Profile



Partial Dependance Profiles (PDP)
Accumulated Local Effects (ALE)

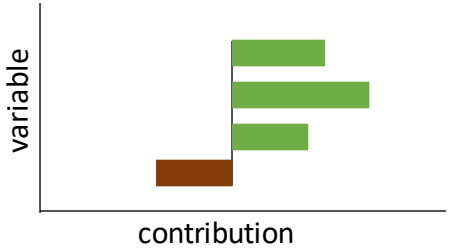
Model Diagnostics



Residual plots
Variable vs. prediction plots
Demographic parity

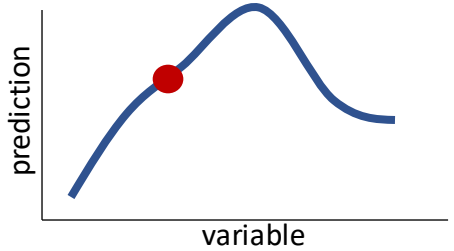
Local Explanations

Predict Parts



Break Down (BD)
SHapley Additive exPlanations (SHAP)
Local Interpretable Model agnostic Explanations (LIME)

Predict Profile

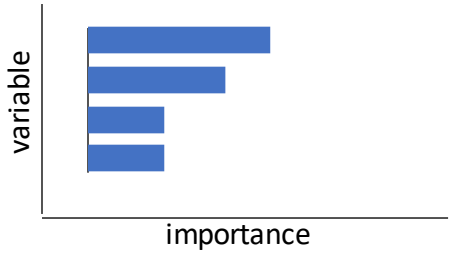


Ceteris Paribus (CP) / Individual Conditional Expectations (ICE)

Taxonomy in terms of explanation object

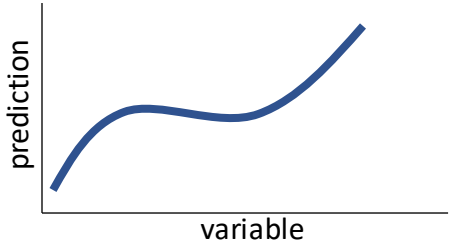
Global Explanations

Model Parts



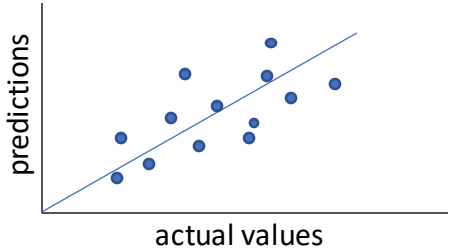
Permutational feature importance
Leave One Covariate Out (LOCO)
Surrogate models
Aggreagated SHapley Additive exPlanations

Model Profile



Partial Dependance Profiles (PDP)
Accumulated Local Effects (ALE)

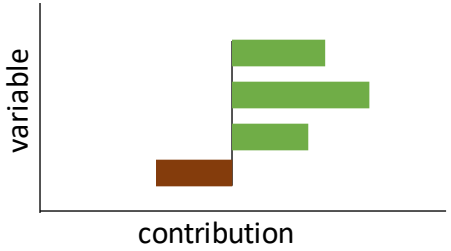
Model Diagnostics



Residual plots
Variable vs. prediction plots
Demographic parity

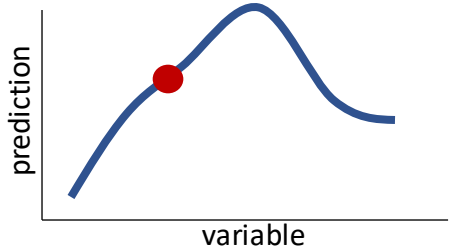
Local Explanations

Predict Parts



Break Down (BD)
SHapley Additive exPlanations (SHAP)
Local Interpretable Model agnostic Explanations (LIME)

Predict Profile



Ceteris Paribus (CP) / Individal Conditional Expectations (ICE)

Predict Diagnostics

Local residual density plot

Six Python libraries for XAI

eli5 (Python)

- One of the oldes XAI libraries.
- Supports the most common Python frameworks and packages: scikit-learn, Keras, xgboost, LightGBM, CatBoost, lightning, and sklearn-crfsuite.
- Good for text, image, and tabular data.

Feature Importance

Weight	Feature
0.4278	Sex=female
0.1949	Pclass=3
0.0665	Embarked=S
0.0510	Pclass=2
0.0420	SibSp
0.0417	Cabin=
0.0385	Embarked=C
0.0358	Ticket=1601
0.0331	Age
0.0323	Fare
0.0220	Pclass=1
0.0143	Parch

eli5 (Python)

LIME

y=alt.atheism (probability **0.000**, score **-9.663**) top features

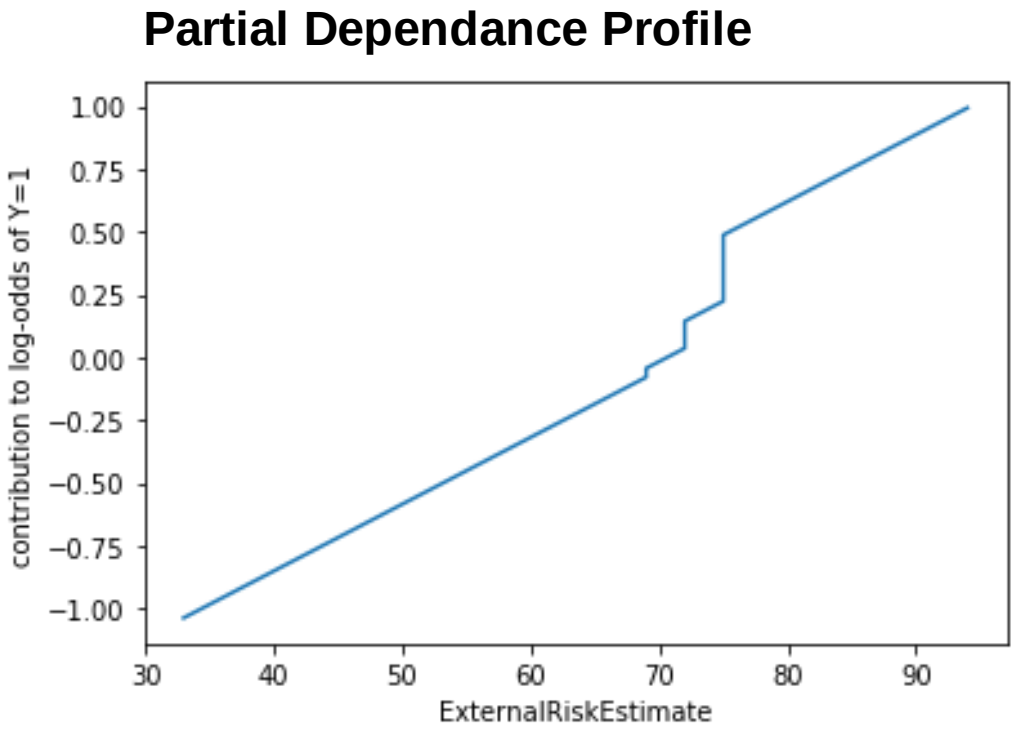
Contribution[?]	Feature
-0.360	<BIAS>
-9.303	Highlighted in text (sum)

as i recall from my bout with kidney stones, there isn't any medication that can do anything about them except relieve the pain. either they pass, or they have to be broken up with sound, or they have to be extracted surgically. when i was in, the x-ray tech happened to mention that she'd had kidney stones and children, and the childbirth hurt less.

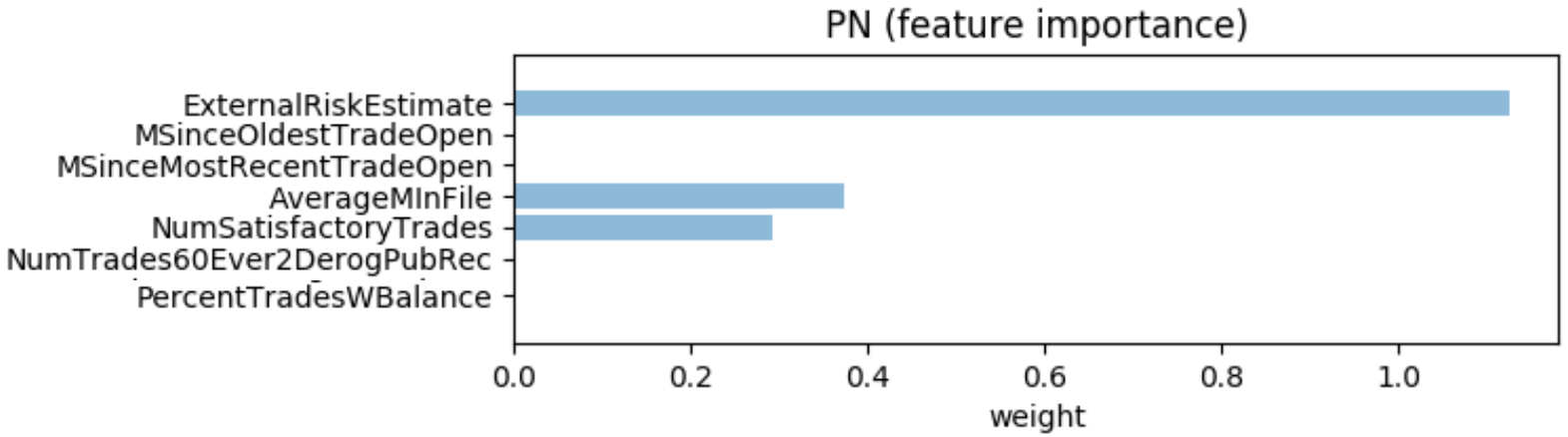
Grad-CAM

aix360 (Python)

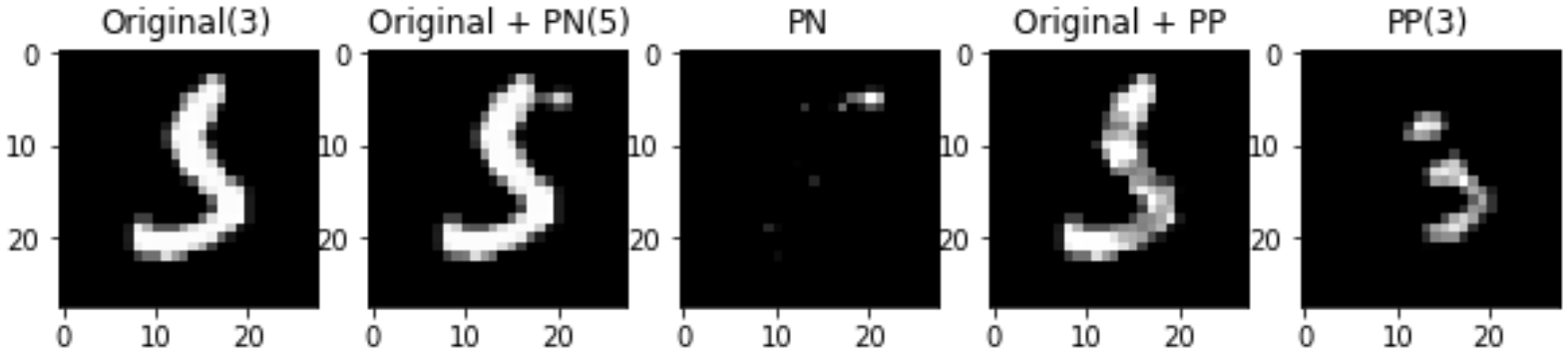
- Explainability of both models and data sets.
- A wide range of methods: local, global, model-specific, model-agnostic.
- Many tutorials and example notebooks



aix360 (Python)

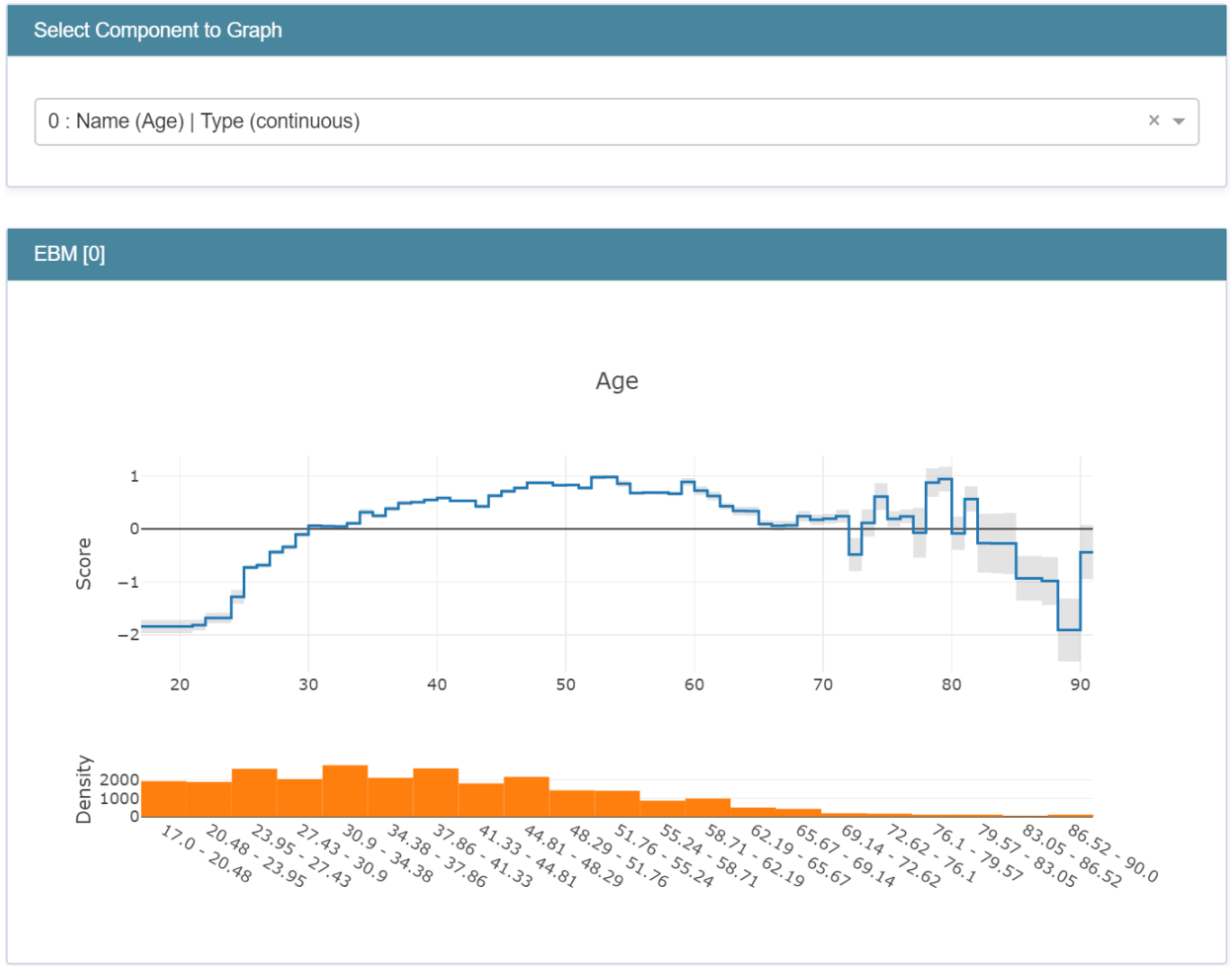


Plot Pertinent Negative (PN) and Pertinent Positive (PP) explanations



interpret (Python)

- Provides interpretable models, such as Explainable Boosting and Decision Rule List.
- Many model-agnostic methods: LIME, SHAP, PDP.
- Provides comparisons of multiple models on interactive dashboards.



Interpret ML Dashboard

Overview | Data | Performance | Global | Local

Select Explanation

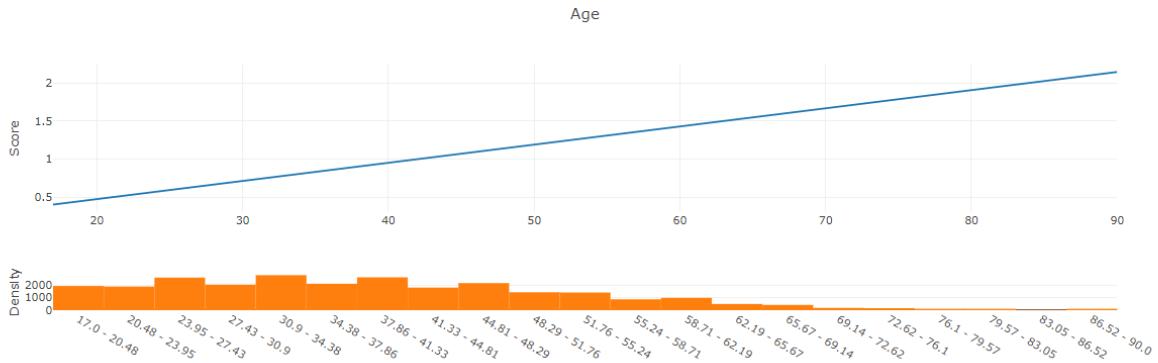
× LR × Tree

Select Components to Graph

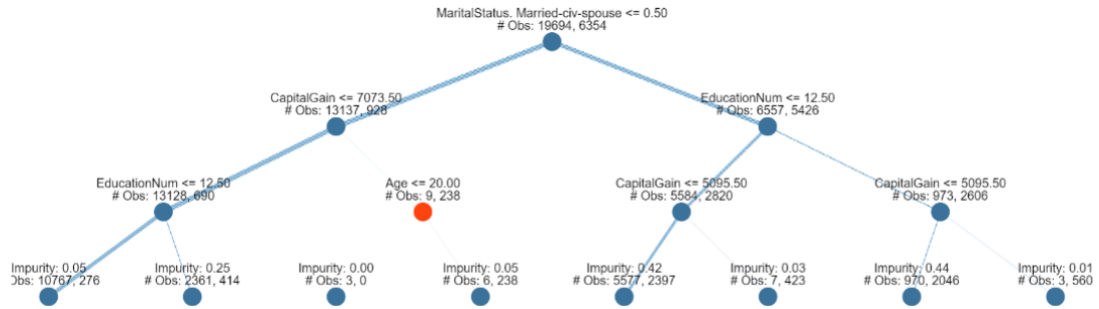
FILTER ROWS

	Name	Type	# Unique	% Non-zero	Importance	SelectID
■	Age	continuous	73	1		0
	fnlwgt	continuous	18385	1		1
	EducationNum	continuous	16	1		2
	CapitalGain	continuous	116	0.084		3
	CapitalLoss	continuous	88	0.047		4
	HoursPerWeek	continuous	93	1		5
	WorkClass. ?	categorical	2	0.056		6

LR [0]



Tree [0]

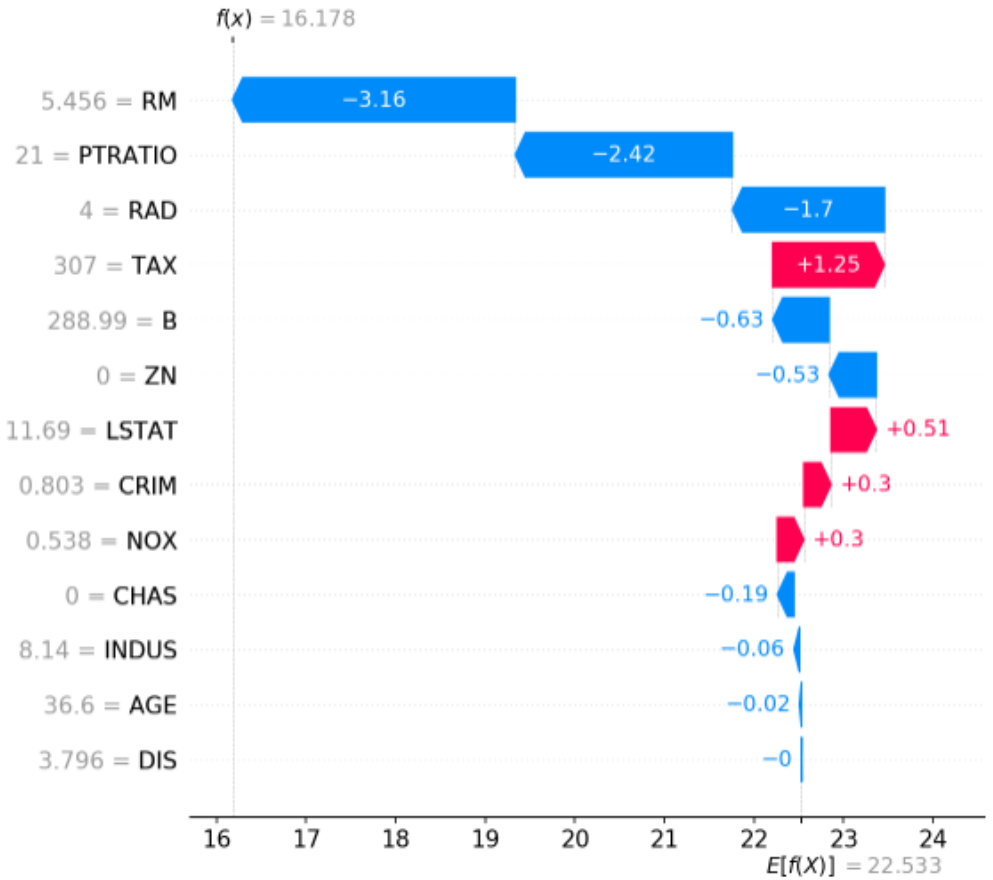


LR (Overall)

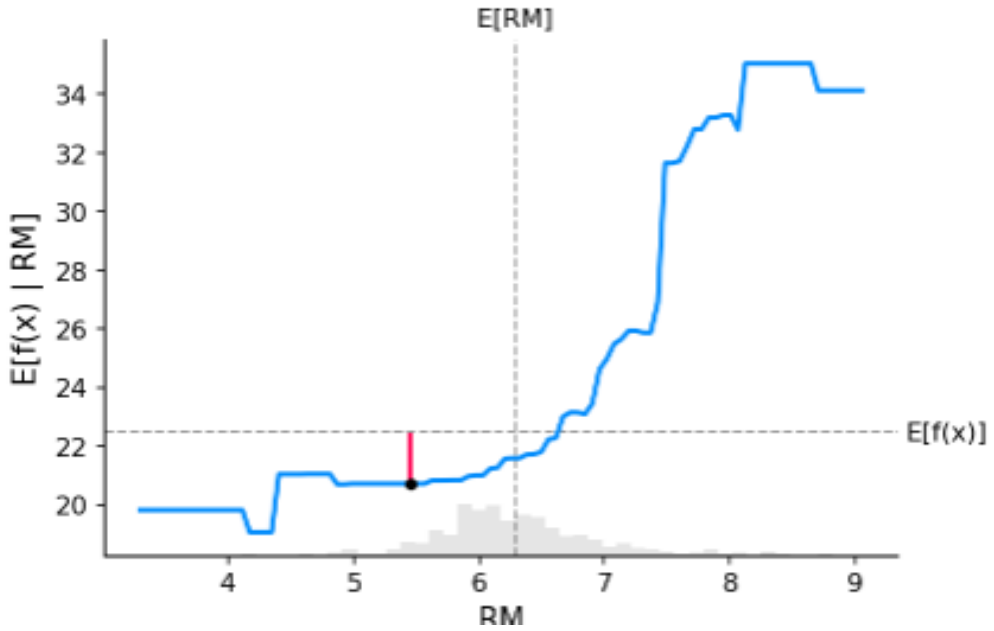
Tree (Overall)

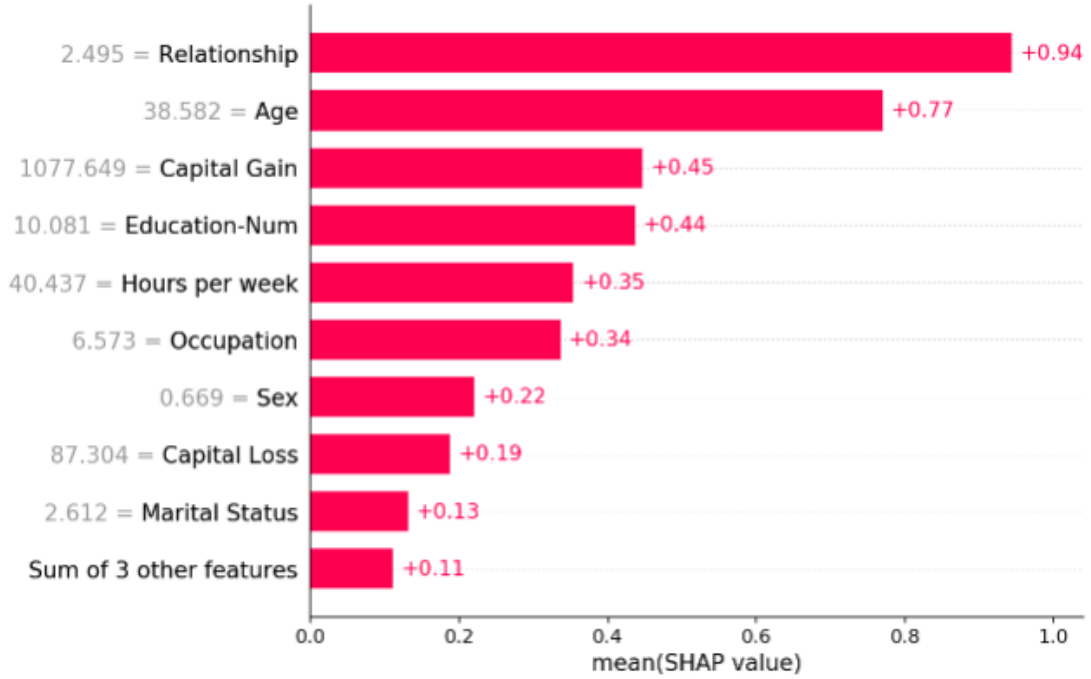
shap (Python)

- Popular and mature
- Strong theoritical background based on game theory
- Model agnostic
- Perfect for tabular image and text data
- Implements methods that comply with different task



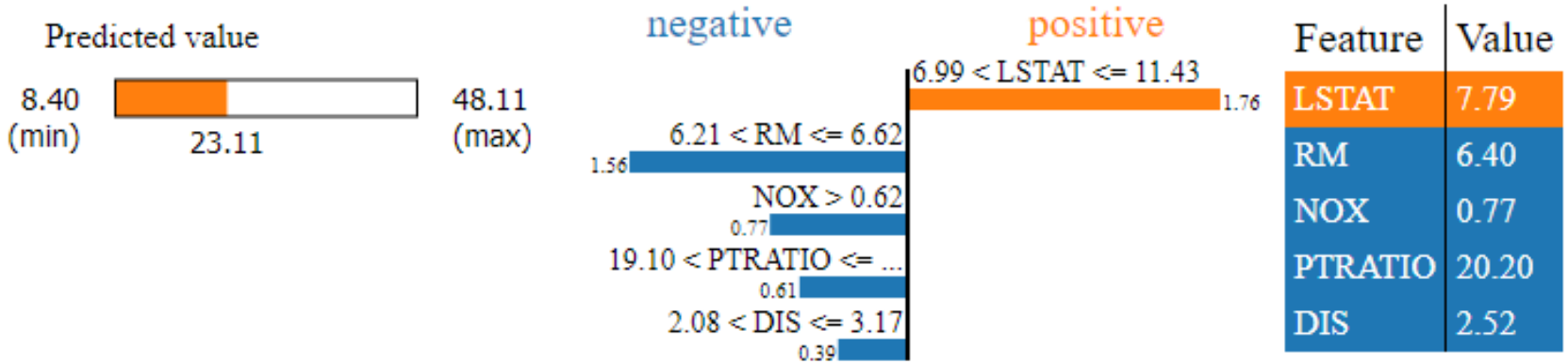
shap (Python)


E[RM]
E[f(x) | RM]
34
32
30
28
26
24
22
20
E[f(x)]
4
5
6
7
8
9
RM

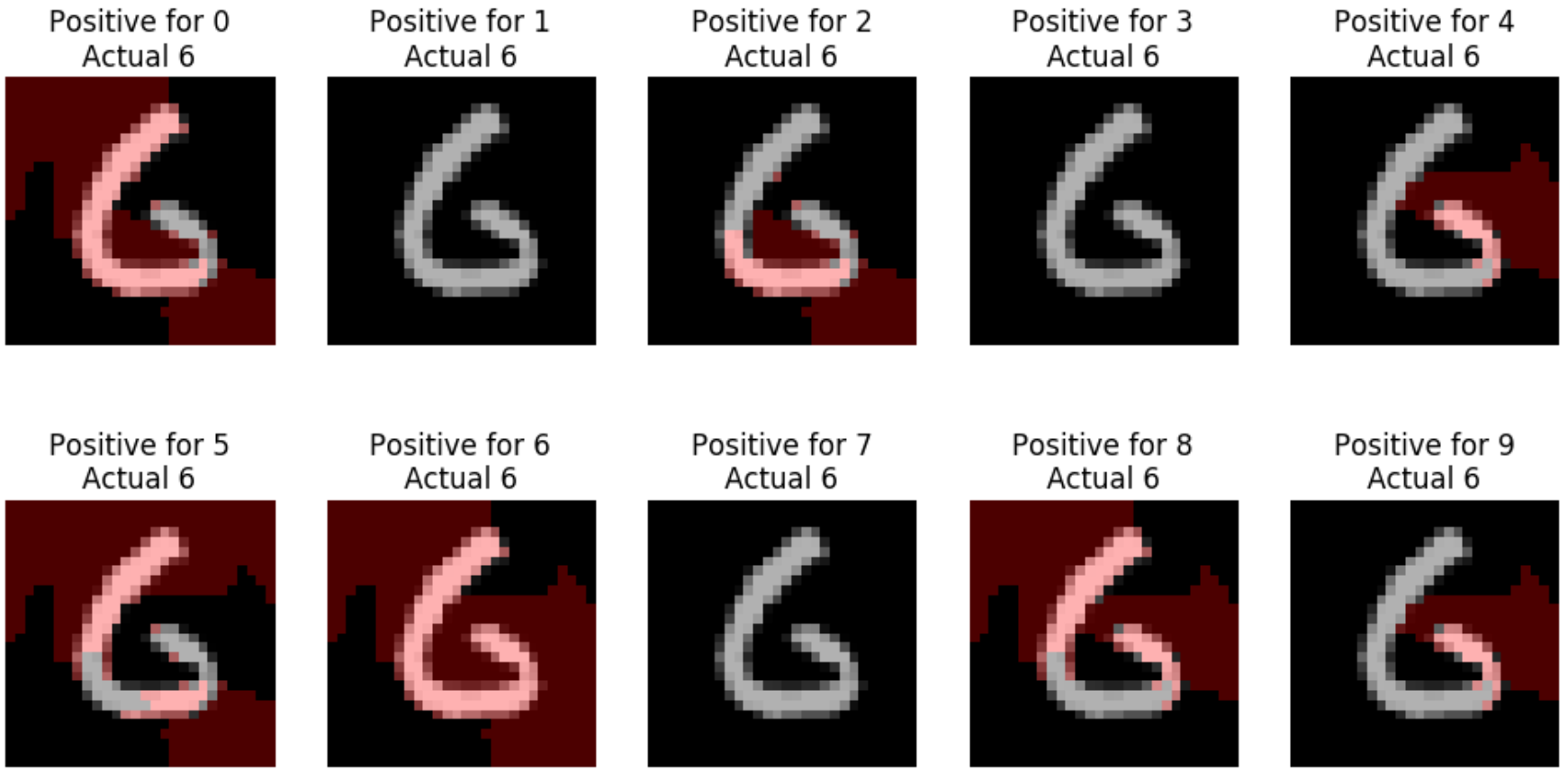

2.495 = Relationship
+0.94
38.582 = Age
+0.77
1077.649 = Capital Gain
+0.45
10.081 = Education-Num
+0.44
40.437 = Hours per week
+0.35
6.573 = Occupation
+0.34
0.669 = Sex
+0.22
87.304 = Capital Loss
+0.19
2.612 = Marital Status
+0.13
Sum of 3 other features
+0.11
0.0
0.2
0.4
0.6
0.8
1.0
mean(SHAP value)

lime (Python)

- Popular and mature
- Perfect for non-tabular data and models with houdndreds of features
- Assumes local linearity of data

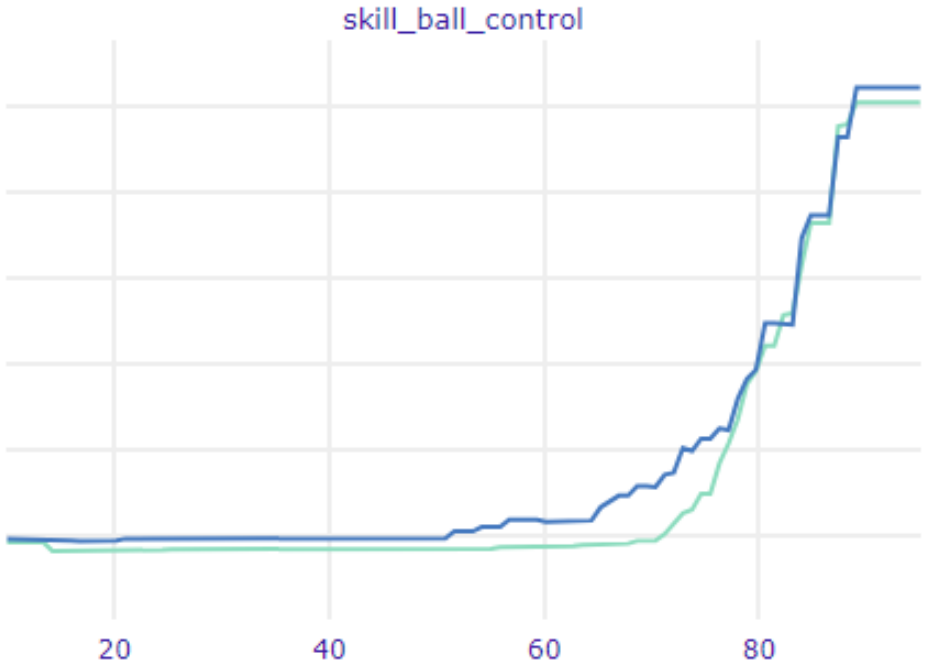


lime (Python)

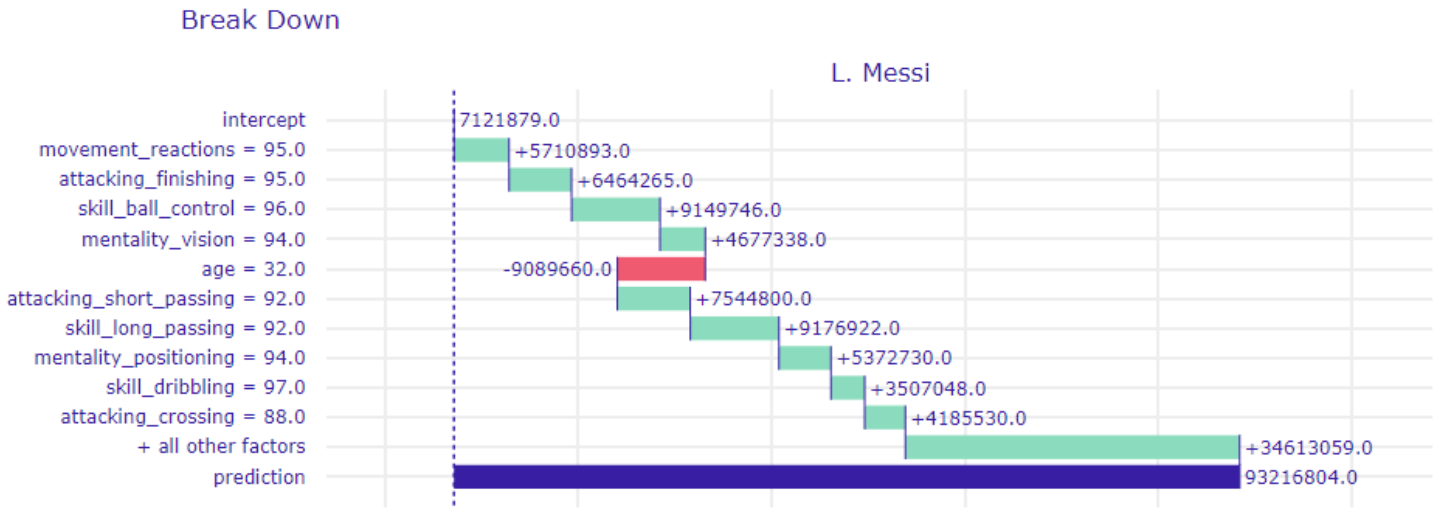

Positive for 0
Actual 6
Positive for 1
Actual 6
Positive for 2
Actual 6
Positive for 3
Actual 6
Positive for 4
Actual 6
Positive for 5
Actual 6
Positive for 6
Actual 6
Positive for 7
Actual 6
Positive for 8
Actual 6
Positive for 9
Actual 6

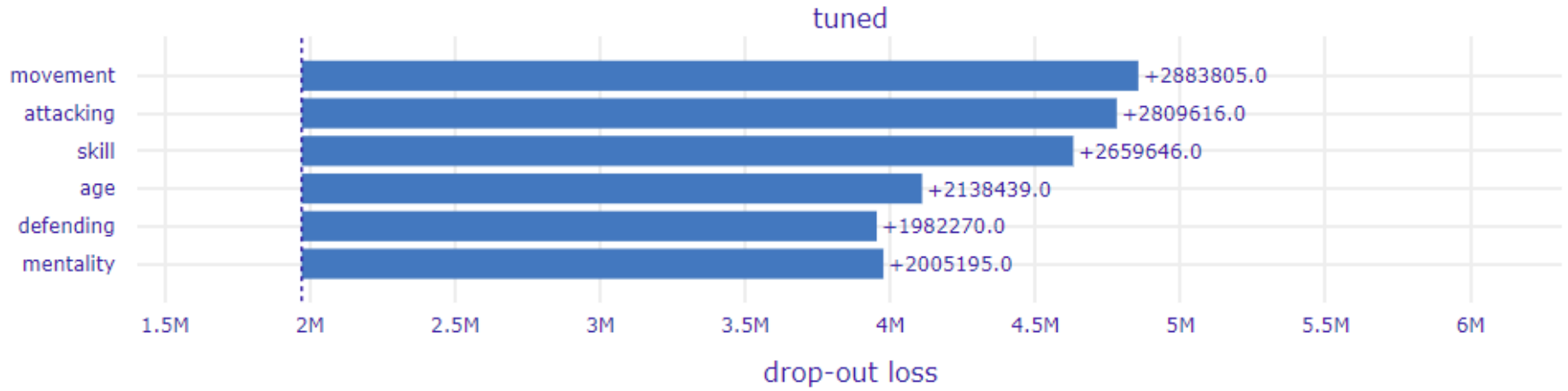
dalex

- Combain library for plenty of explanation methods
- Based on R DALEX package
- Sufficient documentation with theoritical background for all implemented methods



dalex


Break Down
L. Messi
intercept
movement_reactions = 95.0
attacking_finishing = 95.0
skill_ball_control = 96.0
mentality_vision = 94.0
age = 32.0
attacking_short_passing = 92.0
skill_long_passing = 92.0
mentality_positioning = 94.0
skill_dribbling = 97.0
attacking_crossing = 88.0
+ all other factors
prediction
7121879.0
+5710893.0
+6464265.0
+9149746.0
+4677338.0
-9089660.0
+7544800.0
+9176922.0
+5372730.0
+3507048.0
+4185530.0
+34613059.0
93216804.0


tuned
movement
attacking
skill
age
defending
mentality
+2883805.0
+2809616.0
+2659646.0
+2138439.0
+1982270.0
+2005195.0
1.5M
2M
2.5M
3M
3.5M
4M
4.5M
5M
5.5M
6M
drop-out loss

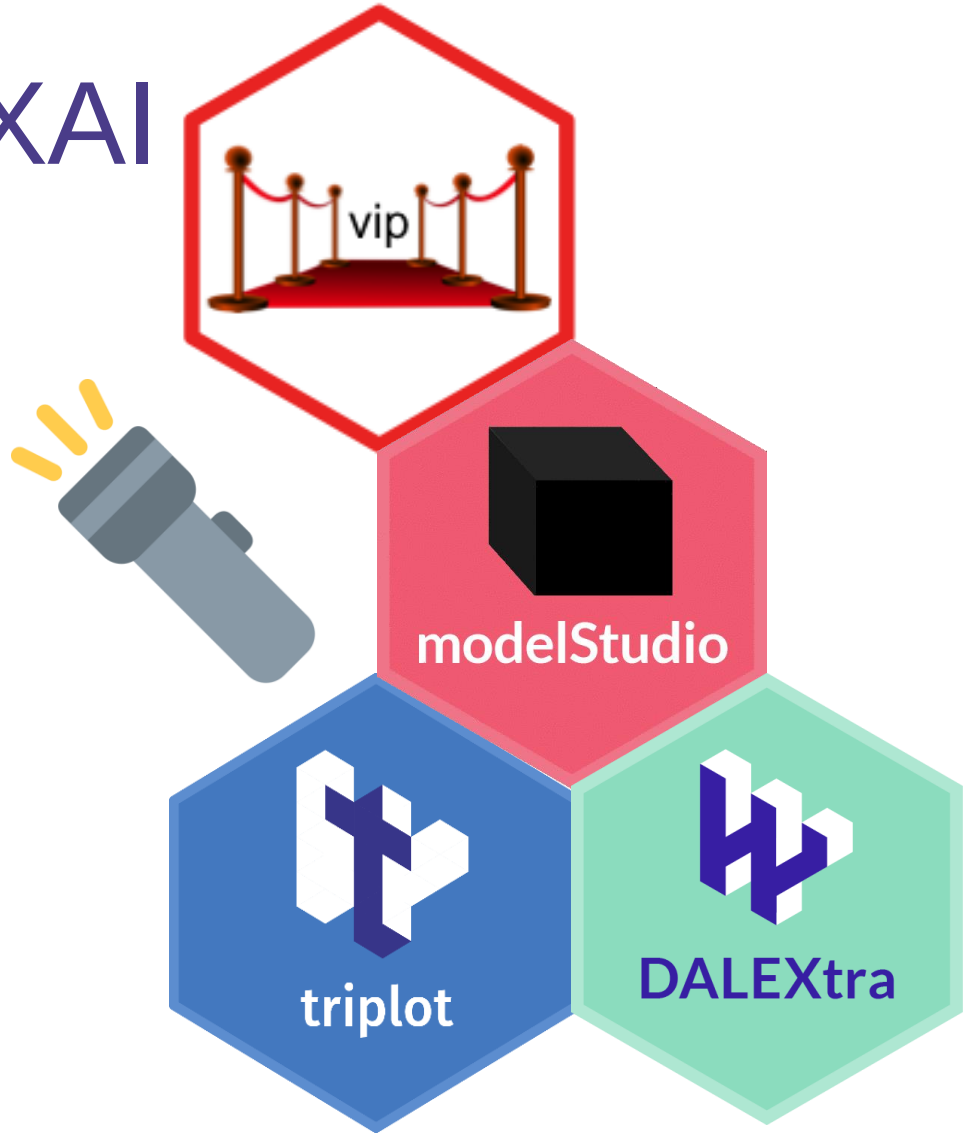
Hot Five R packages for XAI

Hot Five R Packages for XAI

→ Outstand other packages in specific areas.

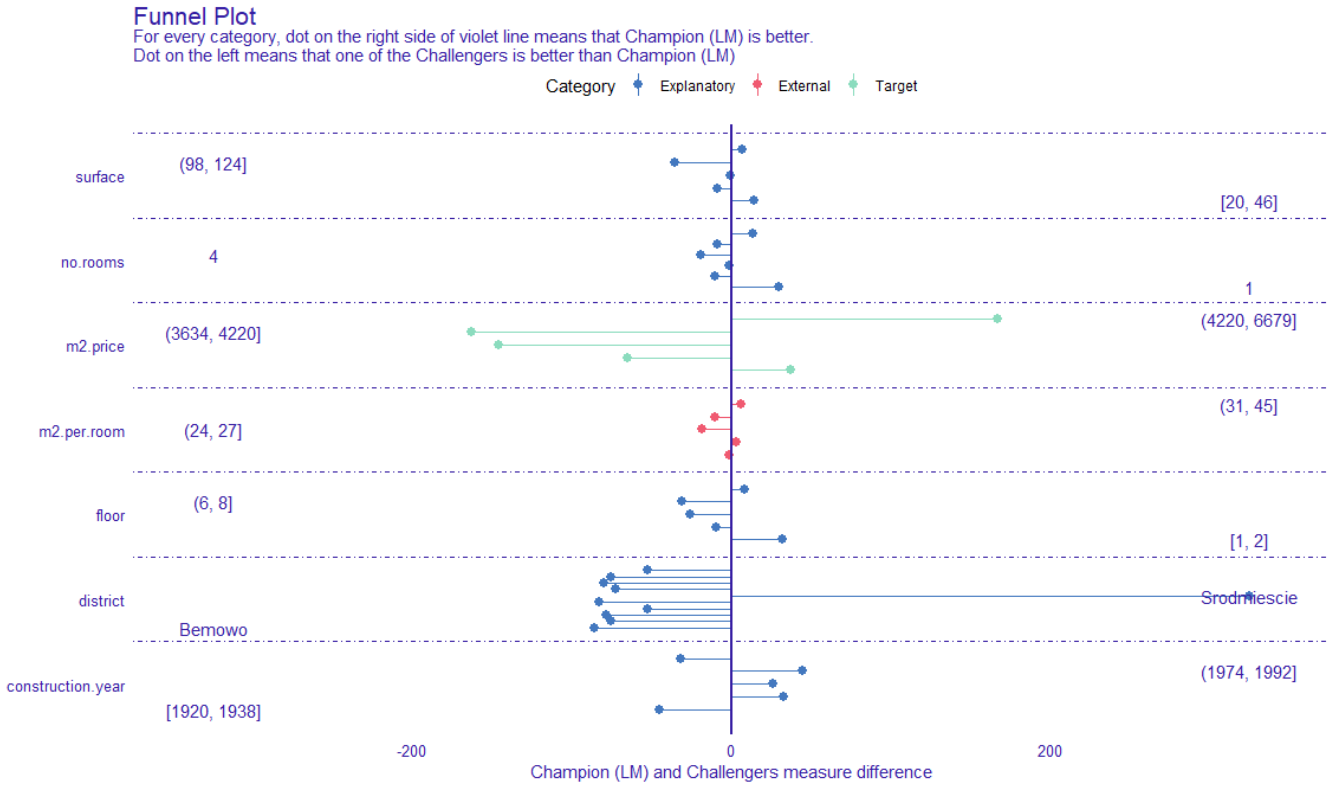
→ The only one that implement interesting conceptions.

→ Answer different XAI issues.



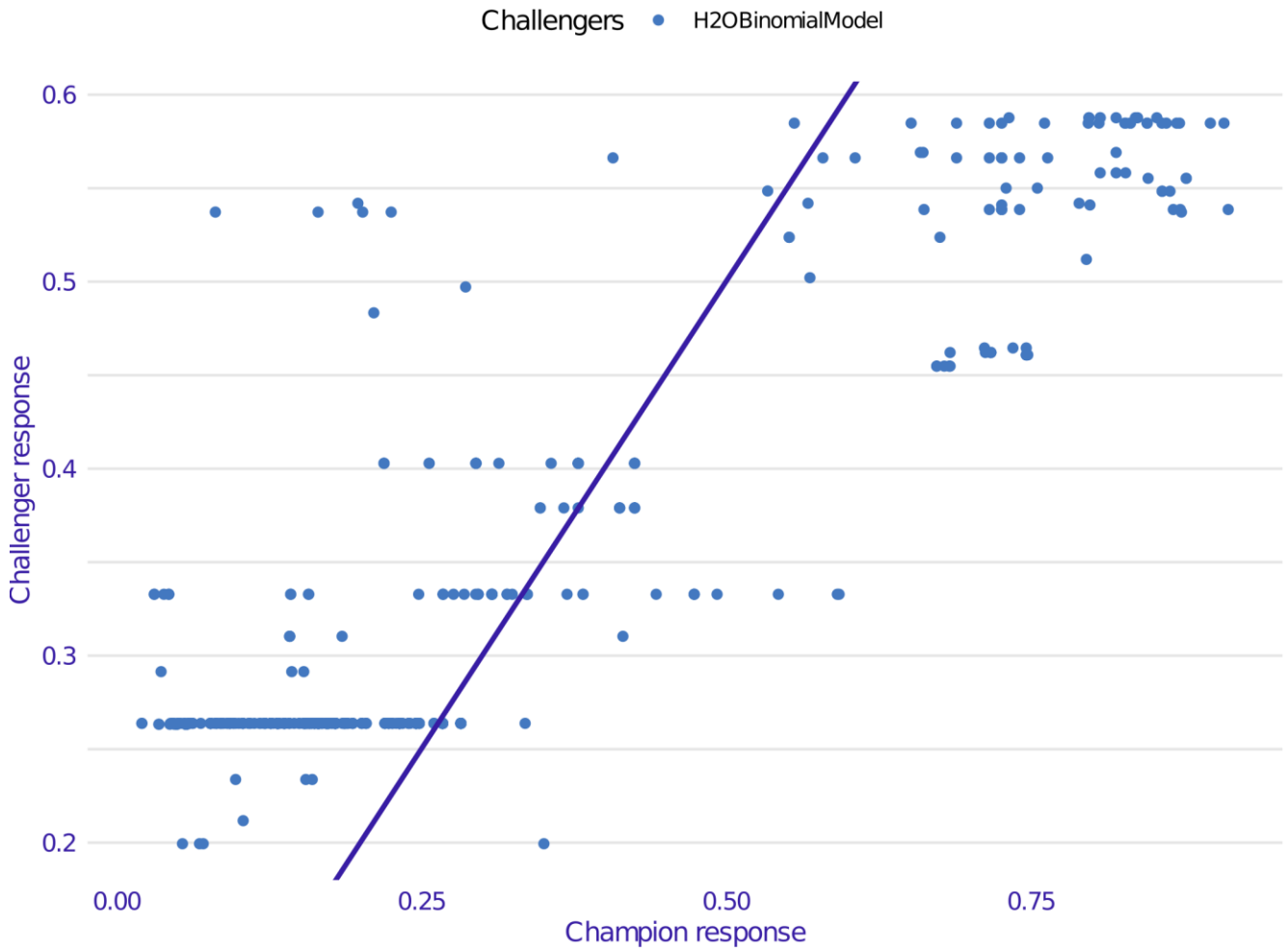
DALEXtra

- Provides various methods for model diagnostics, including Funnel Plot.
- Allows users to comapare more than one predictive model at once.
- Integration with various Machine Learning frameworks, also from different languages.
  For example, mlr, caret, scikit-learn, and more
  Other XAI tools do not!



Use-case on the Titanic dataset is under https://mi2datalab.github.io/XAI-tools/DALEXtra.html.

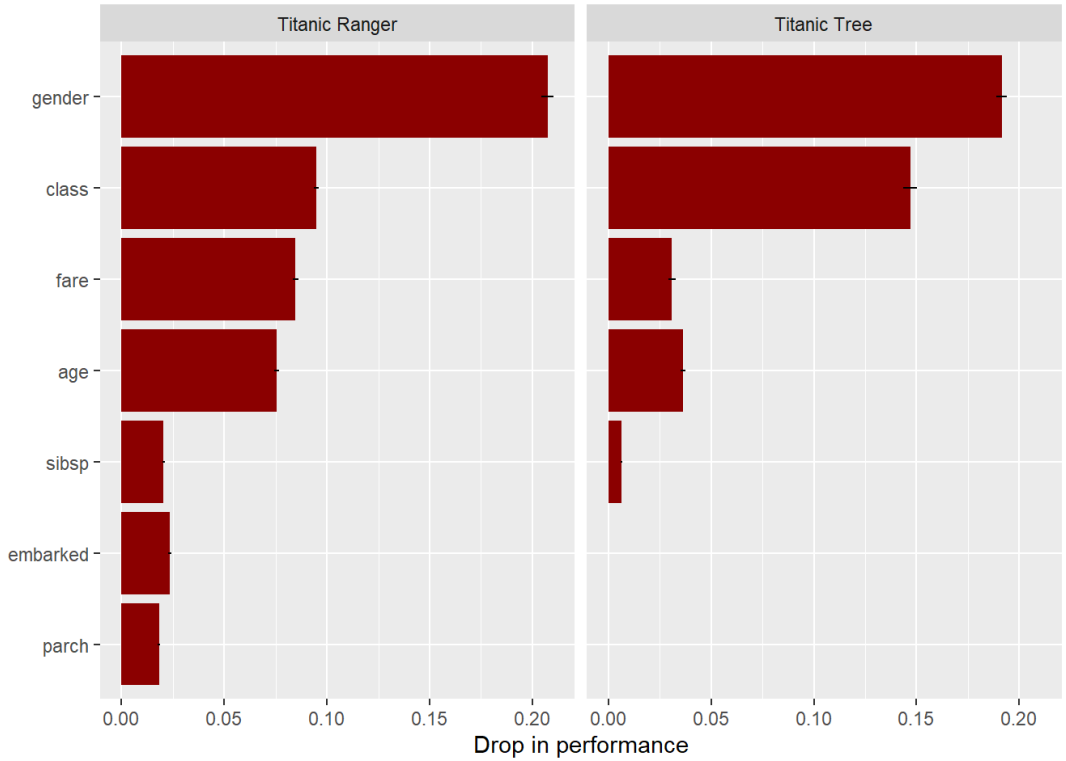
DALEXtra


Challengers
H2OBinomialModel
0.6
0.5
0.4
0.3
0.2
Challenger response
0.00
0.25
0.50
0.75
Champion response

flashlight

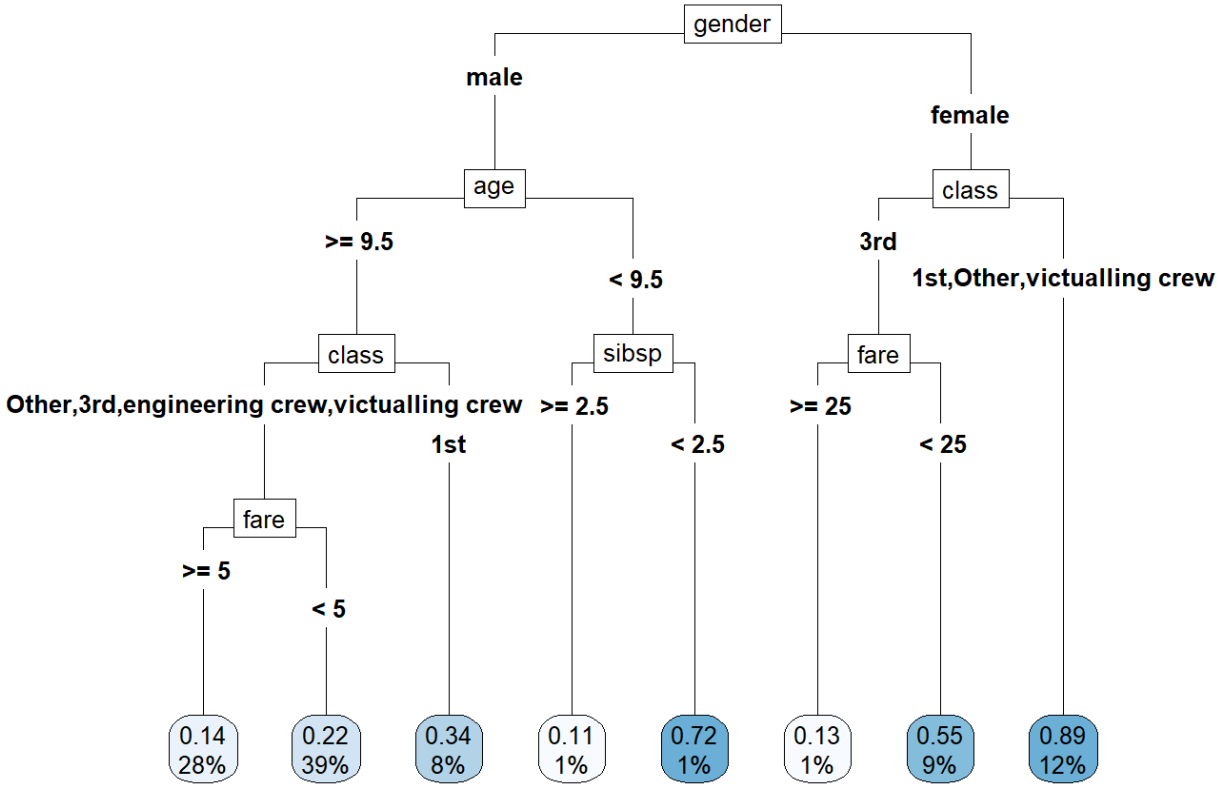
- Provides variable importance, PDP, ALE, residual, target and predicted value profiles. Local explanations: SHAP, BreakDown, and ICE.

- Observations can be weighted and taken into consideration while computing explanations
  No such feature in DALEX-verse or iml!



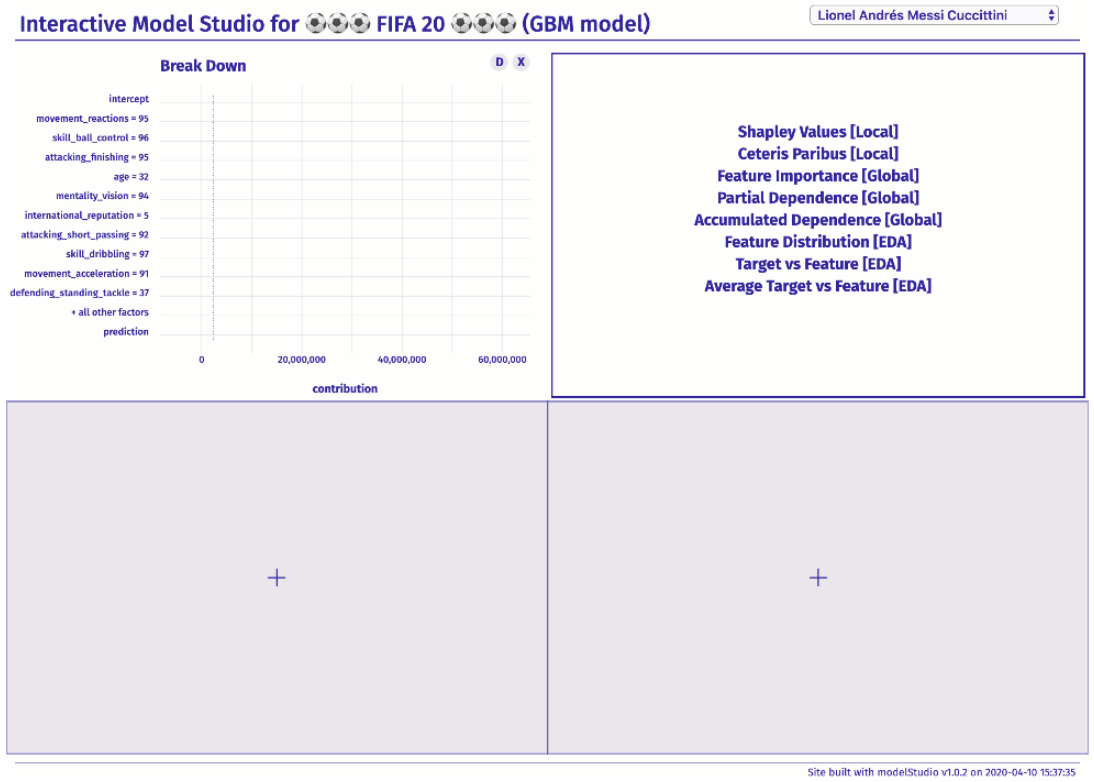
Use-case on the Titanic dataset is under https://mi2datalab.github.io/XAI-tools/flashlight.html.

flashlight


gender
male
female
age
class
>= 9.5
< 9.5
3rd
1st,Other,victualling crew
class
sibsp
fare
Other,3rd,engineering crew,victualling crew
>= 2.5
>= 25
1st
< 2.5
< 25
fare
>= 5
< 5
0.14
28%
0.22
39%
0.34
8%
0.11
1%
0.72
1%
0.13
1%
0.55
9%
0.89
12%

modelStudio

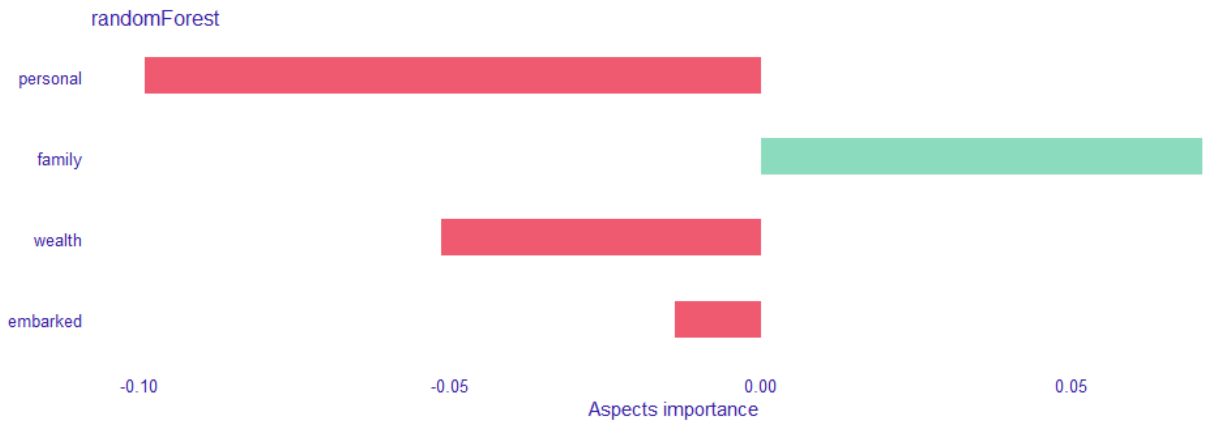
- Flexible stanalone interface to model explanations.
- Ability to smoothly naviagte between many different explanations at the same same time
- Usage does not require fluency in programming.



Use-case on the Titanic dataset is under https://mi2datalab.github.io/XAI-tools/modelStudio_titanic.html

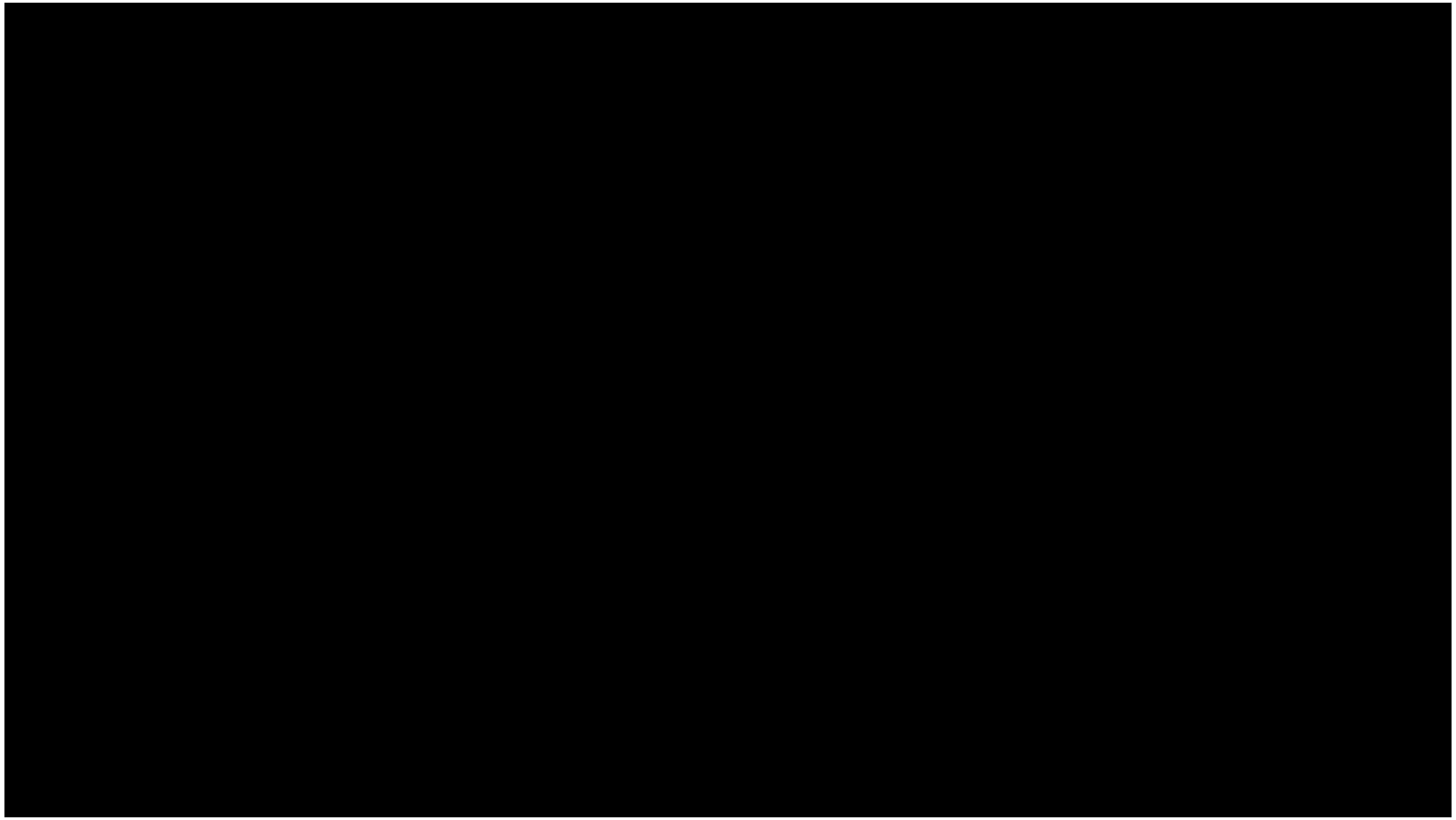
triplot

- A powerful tool that helps explaining models with correlated features
- Allows taking into consideration the whole groups of dependent variables;
- Model agnostic and it works for local and global explanations;



Use-case on the Titanic dataset is under
https://mi2datalab.github.io/XAI-tools/triplot.html

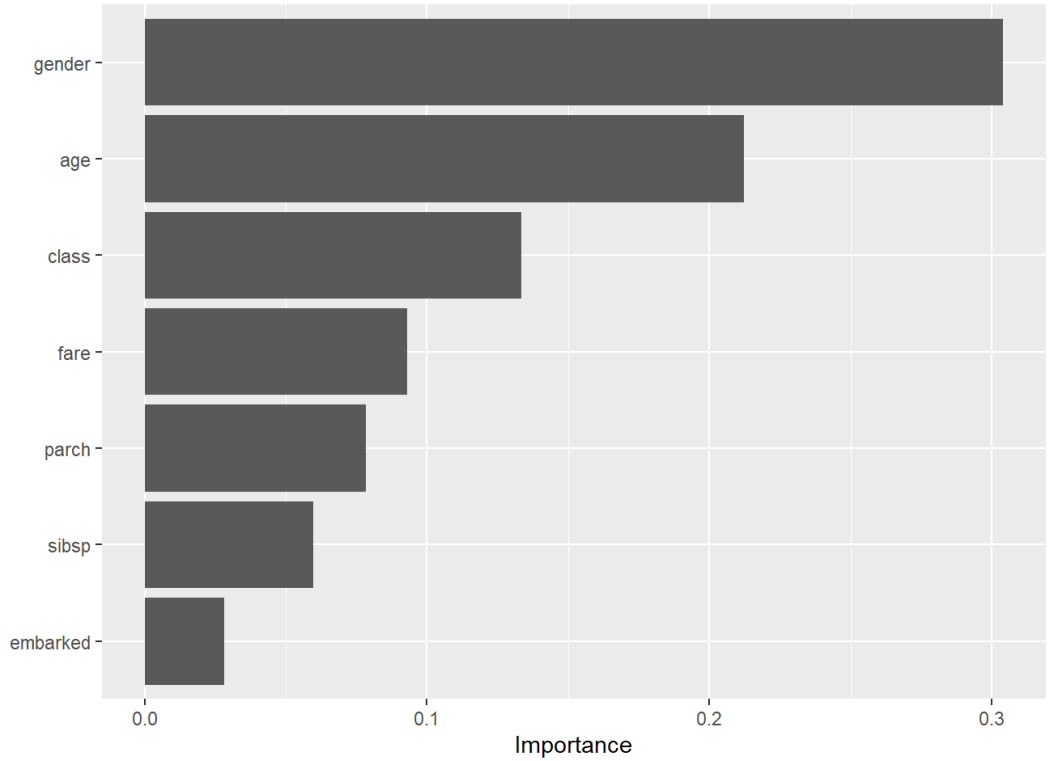
triplot



vip

- Many different ways of computing variable importance, including, permutational feature importance, Shapley-based variable importance, and Variance-based variable importance.

- Offers both, model agnostics and model specific type of explanations, which is unusuall in R packages for XAI.



Use-case on the Titanic dataset is under https://mi2datalab.github.io/XAI-tools/vip.html

Takeouts

- There are a lot of R and Python packages for XAI with various functionalities.
- There is no single tool that will always be the best.
- We presented **just 11**, for more see our preprint:

xai-toolkits.drwhy.ai

You will see comparison of 27 packages and 6 python libraries for XAI. For example:

- What types of explanations the packages contain.
- What ML framework packages are compatible with.
- How to use each package.
- And more ☺

Teaser

		Global Explanations			Local Explanations		
	Package	Model parts	Model profile	Model diagnostics	Predict parts	Predict profile	Predict diagnostics
R	ALEPlot	-	✓	-	-	-	-
	auditor	-	-	✓	-	-	-
	DALEX/DALEXtra	✓	✓	✓	✓	✓	✓
	EIX	✓	-	✓	✓	-	-
	ExplainPrediction	✓	-	-	✓	-	-
	fairness	-	-	✓	-	-	-
	fastshap	✓	✓	-	✓	-	-
	flashlight	✓	✓	-	✓	✓	-
	forestmodel	✓	-	-	-	-	-
	fscaret	✓	-	-	-	-	-
	ICEbox	-	✓	-	-	-	-
	iml	✓	✓	-	✓	✓	-
	lime	-	-	-	✓	-	-
	live	-	-	-	-	✓	-
	mcr	-	-	✓	-	-	-
	modelDown	✓	✓	✓	-	-	-
	modelStudio	✓	✓	-	✓	✓	-
	pdp	✓	✓	-	-	-	-
	randomForestExplainer	✓	-	-	-	-	-
	shapper	-	-	-	✓	-	-
	smbinning	✓	-	✓	-	-	-
	survxai	✓	-	✓	✓	✓	-
	vip	✓	-	-	-	-	-
	vivo	✓	-	-	✓	-	-
Python	aix360	✓	✓	✓	✓	-	-
	eli5	✓	-	-	✓	-	-
	interpret	✓	✓	-	✓	-	-
	lime	-	-	-	✓	-	-
	shap	✓	✓	-	✓	-	-
	skater	✓	✓	-	✓	-	-

		R				Python		Java
	Package	mlr	mlr3	parsnip	caret	keras	scikit-learn	h2o
R	ALEPlot	⋆	⋆	⋆	⋆	•	•	⋆
	auditor	⋆	⋆	✓	✓	•	•	⋆
	DALEX	⋆	⋆	✓	✓	•	•	⋆
	DALEXtra	✓	✓	✓	✓	✓	✓	✓
	EIX[1]	-	-	-	-	-	-	-
	ExplainPrediction	⋆	⋆	⋆	⋆	•	•	⋆
	fairness	⋆	⋆	⋆	⋆	•	•	⋆
	fastshap	⋆	⋆	⋆	⋆	•	•	⋆
	flashlight	⋆	⋆	⋆	⋆	•	•	⋆
	forestmodel[2]	-	-	-	-	-	-	-
	fscaret	-	-	-	✓	-	-	-
	iBreakDown	⋆	⋆	✓	✓	•	•	⋆
	ICEbox	⋆	⋆	⋆	⋆	•	•	⋆
	iml	✓	⋆	⋆	✓	•	•	⋆
	ingredients	⋆	⋆	✓	✓	•	•	⋆
	lime	✓	⋆	✓	✓	•	•	✓
	live	⋆	⋆	⋆	⋆	•	•	⋆
	mcr[3]	-	-	-	-	-	-	-
	modelDown	⋆	⋆	✓	✓	•	•	⋆
	modelStudio	✓	✓	✓	✓	✓	✓	✓
	pdp	⋆	⋆	⋆	⋆	•	•	⋆
	randomForestExplainer[4]	-	-	-	-	-	-	-
	shapper	⋆	⋆	⋆	⋆	•	•	⋆
	smbinning[5]	-	-	-	-	-	-	-
	survxai	⋆	⋆	⋆	-	-	-	-
	vip	⋆	⋆	✓	✓	•	•	⋆
	vivo	⋆	⋆	✓	✓	•	•	⋆
Python	aix360[6]	-	-	-	-	-	✓	-
	eli5	-	-	-	-	✓	✓	-
	interpret[7]	-	-	-	-	-	✓	-
	lime	-	-	-	-	✓	✓	⋆
	shap	-	-	-	-	✓	✓	⋆
	skater	•	•	•	•	✓	✓	⋆

pdp: Partial Dependence Plots

A general framework for constructing partial dependence (i.e., marginal effect) plots from various types machine learning models in R.

lime: Local Interpretable Model-Agnostic Explanations

When building complex models, it is often difficult to explain why the model should be trusted. While global measures such as accuracy are useful, they cannot be used for explaining why a model made a specific prediction. 'lime' (a port of the 'lime' 'Python' package) is a method for explaining the outcome of black box models by fitting a local model around the point in question an perturbations of this point. The approach is described in more detail in the article by Ribeiro et al. (2016) <arXiv:1602.04938>.

DALEX: moDel Agnostic Language for Exploration and eXplanation

Unverified black box model is the path to the failure. Opaqueness leads to distrust. Distrust leads to ignoration. Ignoration leads to rejection. DALEX package xrays any model and helps to explore and explain its behaviour. Machine Learning (ML) models are widely used and have various applications in classification or regression. Models created with boosting, bagging, stacking or similar techniques are often used due to their high performance. But such black-box models usually lack of direct interpretability. DALEX package contains various methods that help to understand the link between input variables and model output. Implemented methods help to explore model on the level of a single instance as well as a level of the whole dataset. All model explainers are model agnostic and can be compared across different models. DALEX package is the cornerstone for 'DrWhy.AI' universe of packages for visual model exploration. Find more details in (Biecek 2018) <arXiv:1806.08915>.