Predictive models locally: Explore, Explain and Debug

Local methods are designed to better understand model behaviour around a single observation.

Prepare model explainer

Models are created in different languages with various libraries. New libraries will emerge, existing libraries will change. And they have different internal model structures.

The DALEX ::explain() function creates a model adapter: uniform interface that can be then used by model explainers.

library("DALEX")
explain(model, data, y, label,
predict_func, residual_fun)

General workflow

Function explain() turns models into *explainers* - wrappers with uniform structure.

Specific functions turn *explainers* into *explanations.*

For *explanations* one can use generic functions: print - prints short summary, plot - created ggplot2 plot, plotD3 - creates a D3 chart based on r2d3 package, describe - creates a text summary for the explanation.

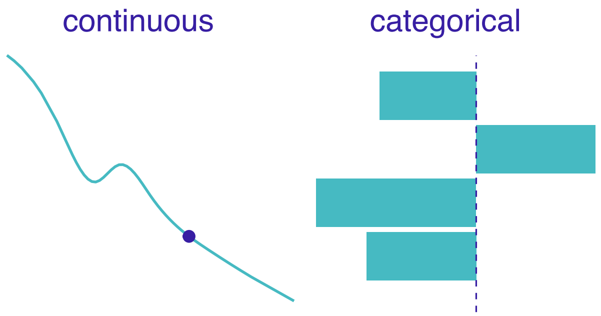
print(explanation)
plot(explanation)
plotD3(explanation)
describe(explanation)

Ceteris Paribus Profiles

How the model response would change for a particular observation if only a single feature is changed?

Best for:
What if questions. Small number of interpretable features.
Be careful when:
Features are correlated.

library("ingredients")
ceteris_paribus(explainer,
observation, variables)

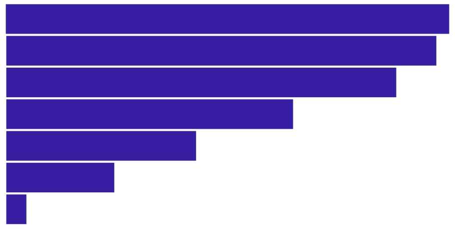


Profile Oscillations

How sensitive is the model response on individual features?

Best for:
Selection interesting CP profiles
Be careful when:
Features are correlated.

calculate_oscillations(explanation)

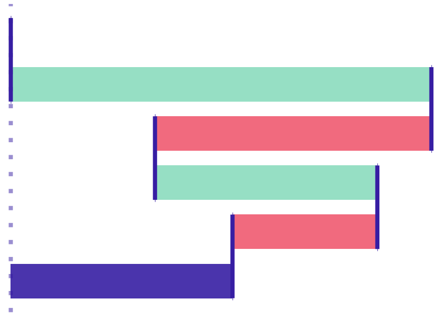


Break Down attributions

How the average model response change when new features are being fixed in the observation of interest?

Best for:
Why questions. Moderate number of features.
Be careful when:
Features are correlated.

library(“iBreakDown")
break_down(explainer, observation)

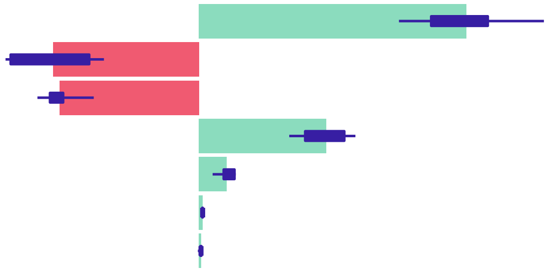


Shapley additive values

How the model response can be decompose into additive attributions.

Best for:
Why questions. Moderate number of features.
Be careful when:
Features are correlated. Model has interactions.

shap(explainer, observation)

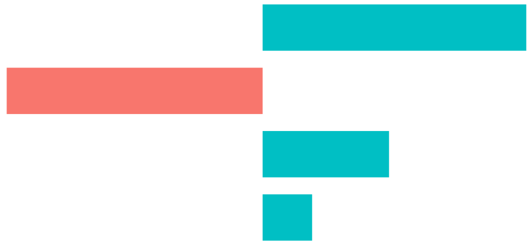


Local Interpretable Model

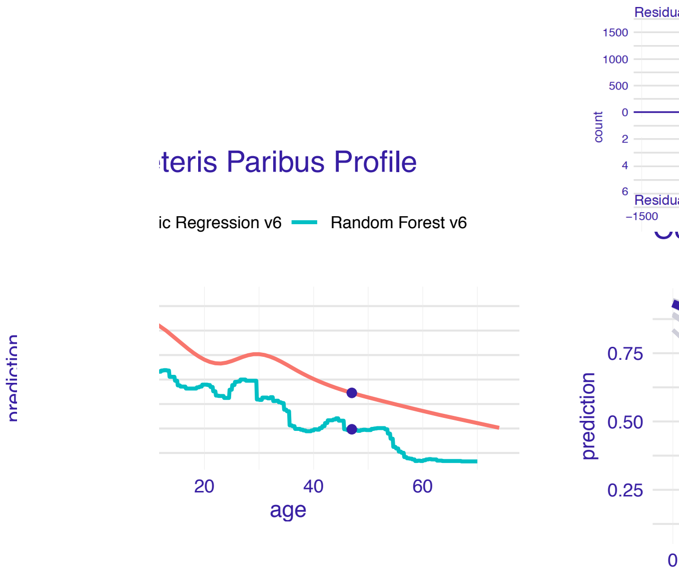
LIME: Local Interpretable Model-Agnostic Explanations. Shows sparse explanations for selected aspects.

Best for:
Why questions. Large number of non-interpretable features.
Be careful when:
Sparse explanations make no sense. It is hard to define aspects.

lime(explainer, observation)



Champion challenger



Learn more at **Predictive Models: Explore, Explain, and Debug**• Przemysław Biecek, Tomasz Burzykowski, https://pbiecek.github.io/PM_VEE/