Predictive models: Explore, Explain and Debug (locally)

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

Prepare model explainer (Ch. 2)

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

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

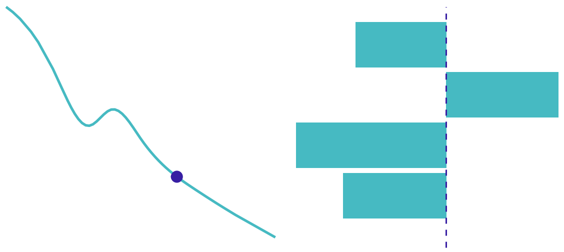
The DALEX ::explain() function creates a model adapter: an uniform interface that can be then used for model exploration and explanations.

Ceteris Paribus Profiles (Ch. 6)

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)
```

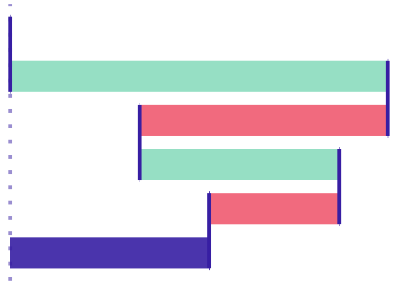


Break Down attributions (Ch. 9)

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)
```



Local Interpretable Model (Ch. 12)

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 have no sense.

```
library(“ingredients")
lime(explainer, observation)
```

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 - short text summary, plot - a ggplot2 plot, plotD3 - a D3 chart based on r2d3 package, describe - a text summary for an explanation.

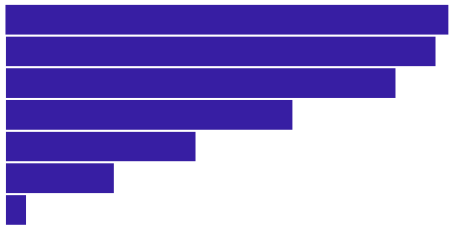
```
print(explanation)
plot(explanation)
plotD3(explanation)
describe(explanation)
```

Profile Oscillations (Ch. 7)

How sensitive is the model response on individual features?

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

```
calculate_oscillations(explanation)
```

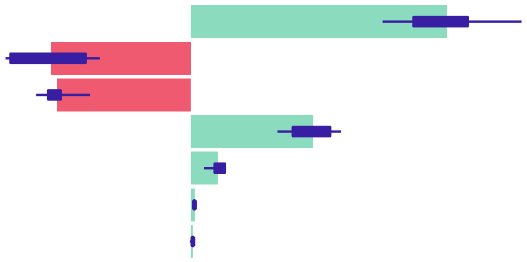


Shapley additive values (Ch. 11)

How the model response can be decompose into additive attributions.

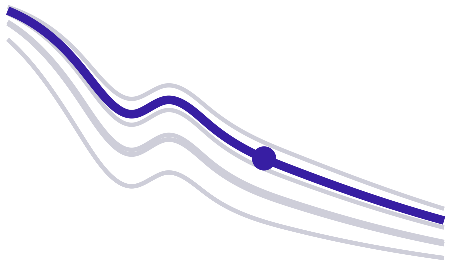
Best for:
 ‘Why’ questions. Moderate number of features.
Be careful when:
 Features are correlated. Model has interactions.

```
shap(explainer, observation)
```

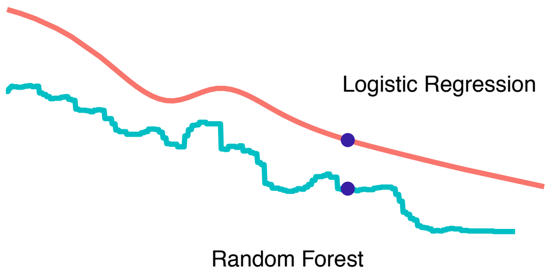


Local diagnostics (Ch. 8)

Instance level analysis of local fit, neighbours ’s residuals and stability.



Two or more explanations can be superimposed on a single plot.



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