Predictive models: Explore, Explain and Debug (locally)

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

Prepare model explainer (Ch. 3)

The DALEX ::explain() function creates model adapters: objects with standardised structure that are used by other methods for model exploration and explanations.

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

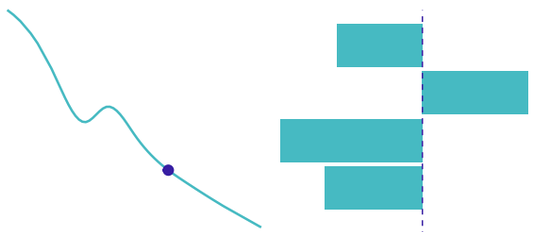
Models can be trained in different languages with various libraries. New libraries will emerge, existing libraries will change. Various models have different structures. This is why we need uniform adapters.

Ceteris Paribus Profiles (Ch. 11)

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

Best for:
What if questions.
A few interpretable features.
Be careful when:
Features are correlated.

```
predict_profile(explainer,
                observation, variables)
```

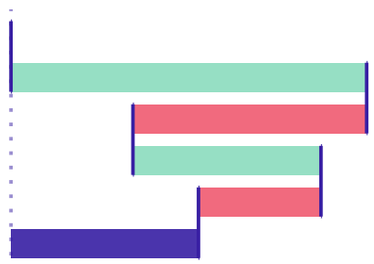


Break Down attributions (Ch. 7)

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.

```
predict_parts(explainer, observation,
        type = "break_down")
```



Local Interpretable Model (Ch. 10)

Local Interpretable Model-Agnostic Explanations (LIME) shows sparse explanations for selected aspects.

Best for:
Why questions.
Lots of non-interpretable features.
Be careful when:
Sparse explanations make no sense.

```
library("DALEXtra")
predict_surrogate(explainer,
        observation, type = "lime")
```

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 plot based on r2d3 package, describe - a text summary for an explanation.

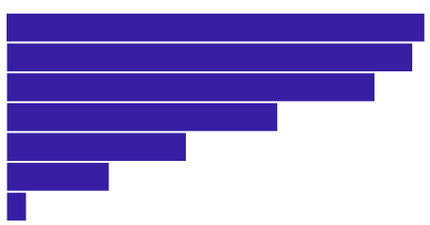
```
print(explanation)
plot(explanation)
plotD3(explanation)
describe(explanation)
```

Profile Oscillations (Ch. 12)

How sensitive is the model response on individual features?

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

```
predict_parts(explainer,
    observation, type = "oscillations")
```

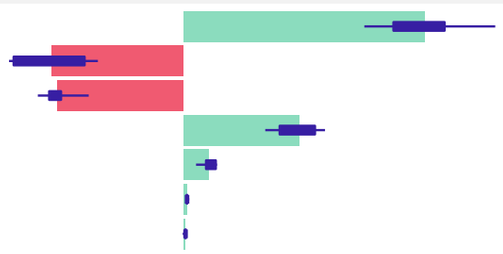


Shapley additive values (Ch. 9)

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.

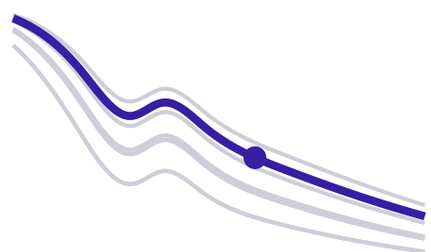
```
predict_parts(explainer, observation,
                type = "shap")
```



Local diagnostics (Ch. 13)

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

Instance level analysis of local fit.
Diagnostic for local residuals.
Stability of predictions.



Learn more at **Explanatory Model Analysis: Explore, Explain, and Debug** Przemysław Biecek, Tomasz Burzykowski, https://pbiecek.github.io/ema/

Unified Model Exploration Process