Imperial College London

Explainability

- The right to an explanation
- Classical Interpretability and Partial Dependence Plots
- **An overview of XAI techniques**
- Are all explanations causal?

Imperial College
London

A classification of XAI algorithms

	Model-agnostic	Model-specific
Global		
Local		

- **Model-agnostic (wrapper) vs model-specific (inline)**: wrapper methods query the model's "predict" API without looking at its internals. Inline methods consider the model internals and therefore are model-specific and do not generalize across model classes.
- **Global vs local:** global methods explain the model's predictive rules at a global level, e.g., feature importance, or global decision rules. Local rules instead attempt to offer explanations for predictions on specific examples.

A classification of XAI algorithms

	Model-agnostic	Model-specific
Global		
Local	Counterfactual explanations	

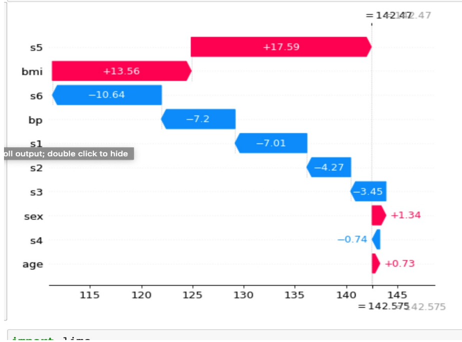
- **Model-agnostic (wrapper) vs model-specific (inline)**: wrapper methods query the model's "predict" API without looking at its internals. Inline methods consider the model internals and therefore are model-specific and do not generalize across model classes.
- **Global vs local:** global methods explain the model's predictive rules at a global level, e.g., feature importance, or global decision rules. Local rules instead attempt to offer explanations for predictions on specific examples.

Quick glance: Shapley values

$$
\begin{aligned}
\mathrm{SHAP}_{\mathrm{BMI}}(30) = w_1\,(f(\mathrm{BMI}=30,\mathrm{age},\mathrm{gender}) - f(\mathrm{age},\mathrm{gender}))\\
+ w_2\,(f(\mathrm{BMI}=30,\mathrm{age}) - f(\mathrm{age}))\\
+ w_3\,(f(\mathrm{BMI}=30,\mathrm{gender}) - f(\mathrm{gender}))\\
+ \dots
\end{aligned}
$$

- SHAP values will be covered in detail in later parts of the course
- Vertical dispersion illustrates presence of interaction effects
- Aggregating SHAP values also gives a feature importance waterfall with sign/directionality.

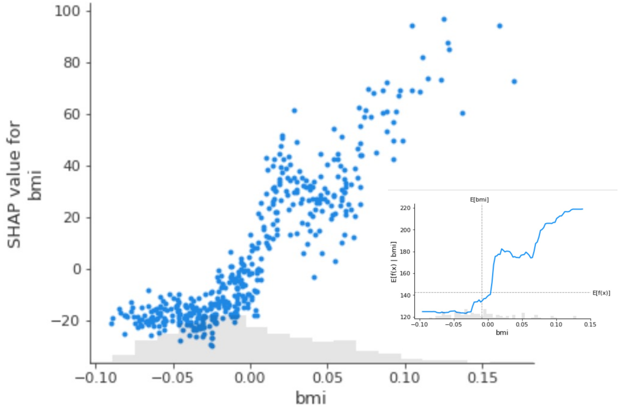
```
shap.plots.waterfall(shap_values_rf[100], max_display=14)
```




```
# explain the RF model with SHAP
explainer_rf = shap.Explainer(rf.predict, X100)
shap_values_rf = explainer_rf(X)
```

```
Exact explainer: 443it [03:31,  2.09it/s]
```

```
shap.plots.scatter(shap_values_rf[:,"bmi"])
```

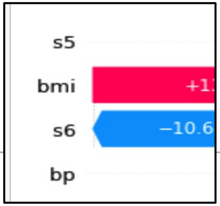


Quick glance: permutation

- Removing the feature altogether (as in SHAP) can create challenges in comparing model predictions
- Instead, we could “destroy” the signal in the feature by permuting its values (i.e., shuffle them).
- We then compare the overall accuracy of the model with or without "shuffling” each feature.
- Technique was popularized by random forests (feature importance)

```
from sklearn.inspection import permutation_importance
r = permutation_importance(rf, X, y, n_repeats=30, random_state=0)
ordered_index = r.importances_mean.argsort()[::-1]
for i in ordered_index:
    print(X.columns[i], "=", r.importances_mean[i].round(4))
```

```
bmi = 0.4809
s5 = 0.4709
bp = 0.1344
s6 = 0.0905
age = 0.0755
s3 = 0.0751
s2 = 0.0659
s1 = 0.0528
s4 = 0.0299
sex = 0.0233
```



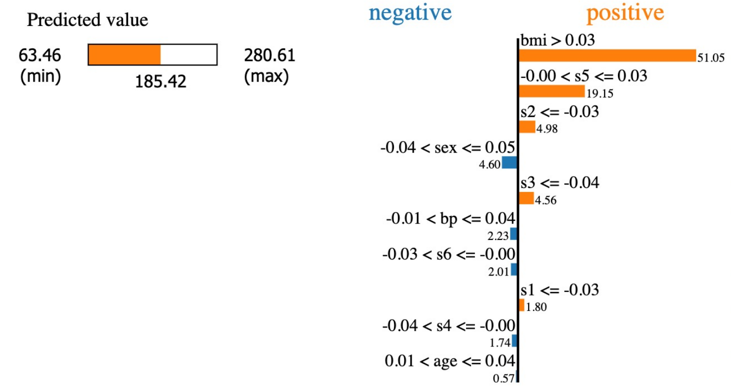
Quick glance: LIME

- LIME fits a local model on a dataset using feature vectors from the neighborhood of the example of interest and their model predictions.
- To allow for some non-linearity, the range of each variable is split into intervals, and treated as categorical.
- Then each feature value in the given example is scored as to whether it pushes the prediction up (positive) or down (negative) holding other things fixed (similar to counterfactuals).

```python
import lime
from lime import lime_tabular

explainer = lime_tabular.LimeTabularExplainer(
    training_data=np.array(X),
    feature_names=X.columns,
    mode='regression'
)
```

```python
exp = explainer.explain_instance(
    data_row=X.iloc[0],
    predict_fn=rf.predict
)
exp.show_in_notebook(show_table=True)
```

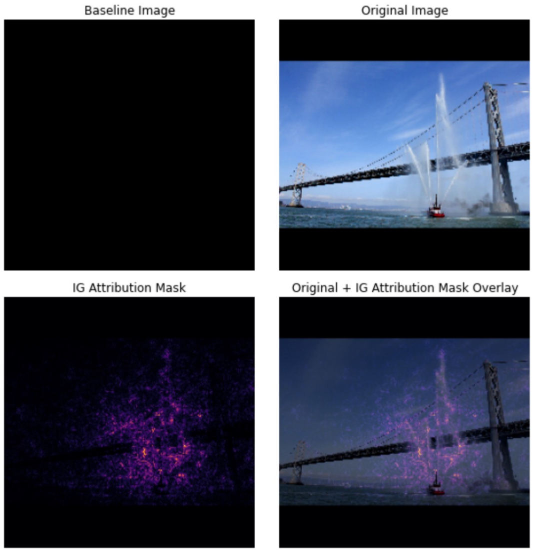


Feature	Value
bmi	0.06
s5	0.02
s2	-0.03
sex	0.05
s3	-0.04
bp	0.02
s6	-0.02
s1	-0.04
s4	-0.00

Imperial College London

Explainability for images

- Some of the above techniques such as LIME also apply to deep learning on images.
- One method in particular is Integrated Gradients which varies an image from a baseline (all-black) to the final image, and computes which pixels have the steepest local slope with respect to the output.



Imperial College London

Summary

- Explainability can be classified as model agnostic (which usually takes place after the model has been trained) or model specific (which can occur inline as a restriction of the model class); as well as local (referring to specific example) or global (attempting to explain the model across examples)
- Common approaches include Shapley values, LIME, Permutation factors and more advanced methods that are well suited to explaining models on non-tabular data, in particular images.
- Overall, feature importance, and counterfactuals, are core “capabilities” of such methods.