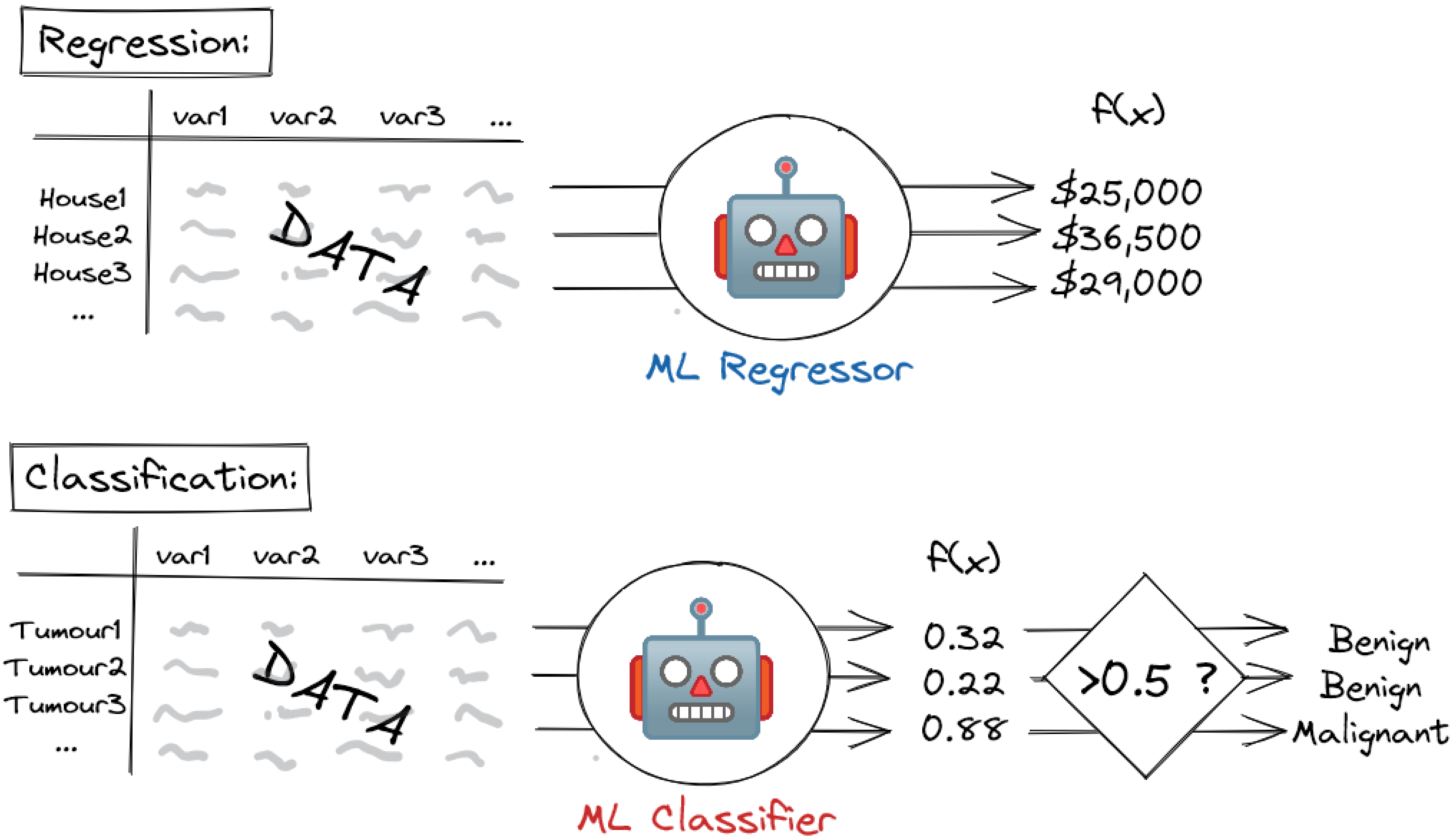
SHAP Analysis Extra Note


Regression:
var1
var2
var3
...
House1
House2
House3
...
DATA
ML Regressor
f(x)
$25,000
$36,500
$29,000
Classification:
var1
var2
var3
...
Tumour1
Tumour2
Tumour3
...
DATA
ML Classifier
f(x)
0.32
0.22
0.88
>0.5 ?
Benign
Benign
Malignant

SHAP

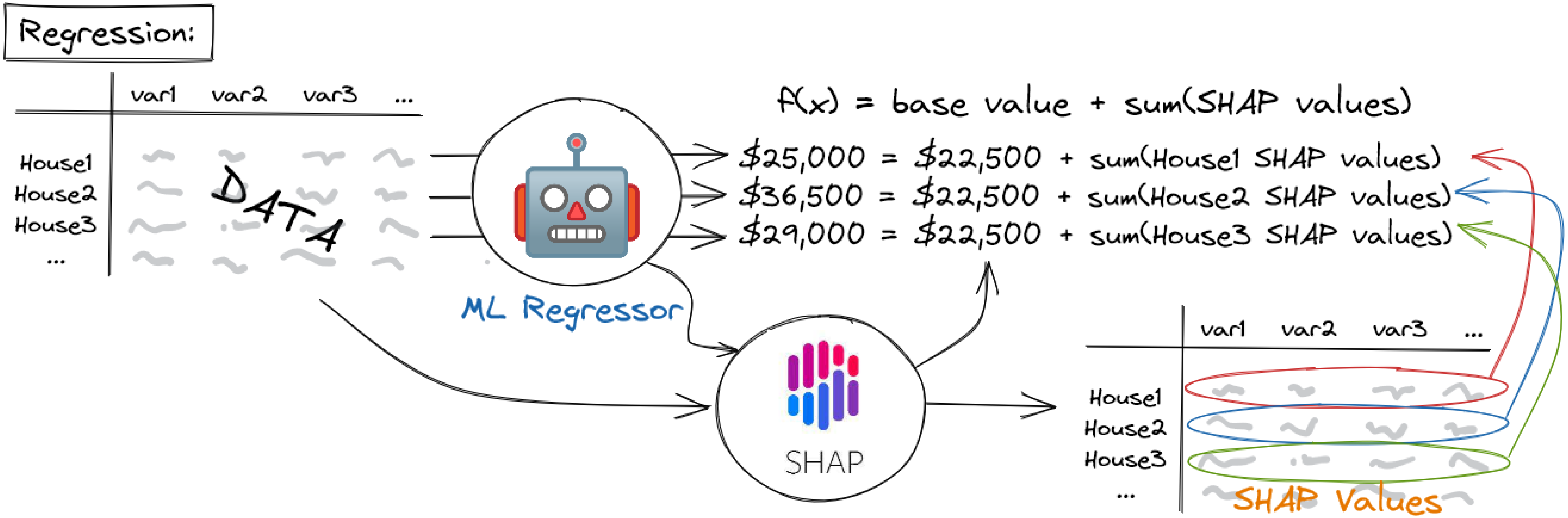
- SHAP = *SHapley Additive exPlanations*

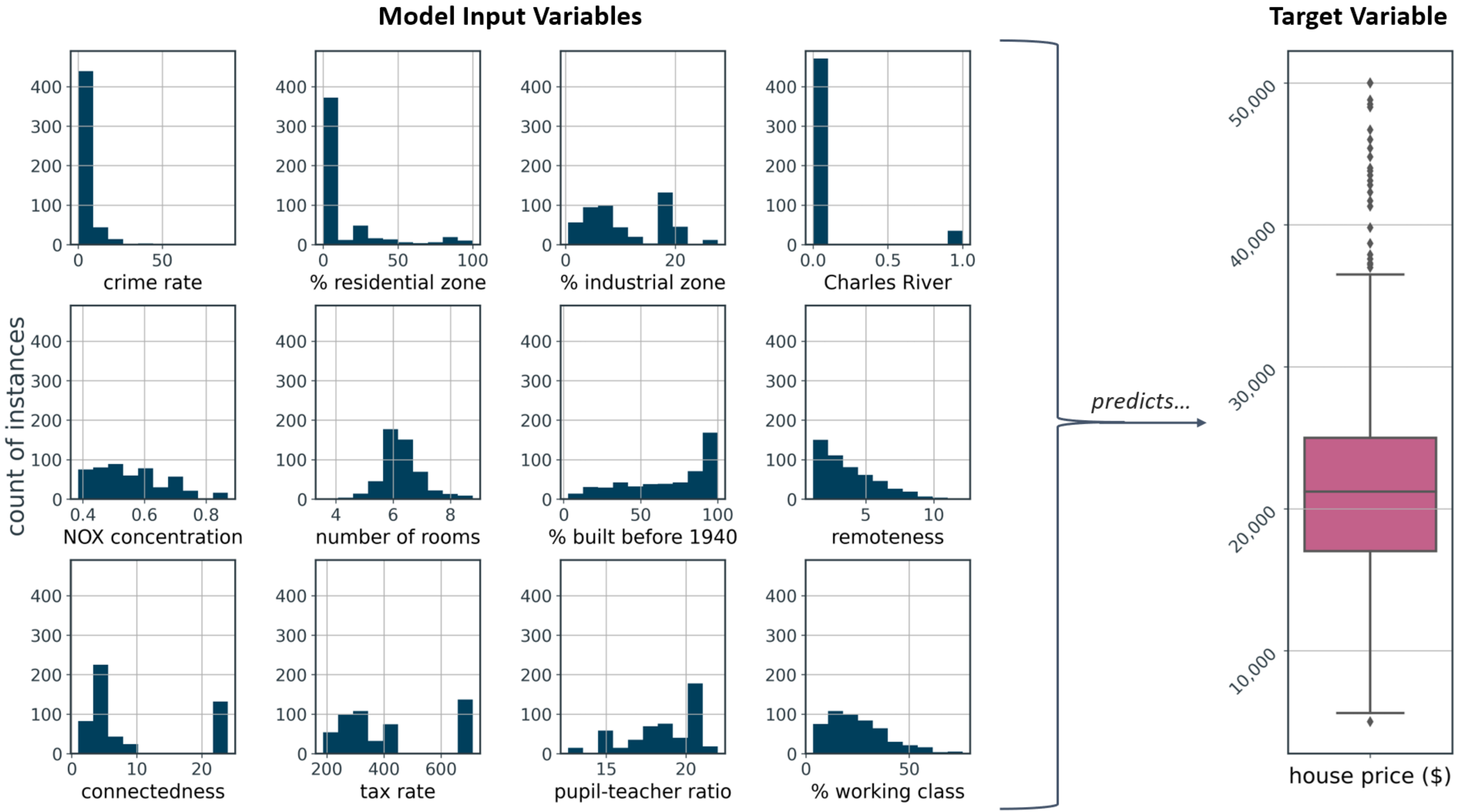
- Popularized use of Shapley values in ML
  - Also used in earlier work by Lipovetsky & Conklin (2001), Strumbelj et al. (2009), Datta et al. (2016)

- SHAP uses Shapley values to explain individual predictions

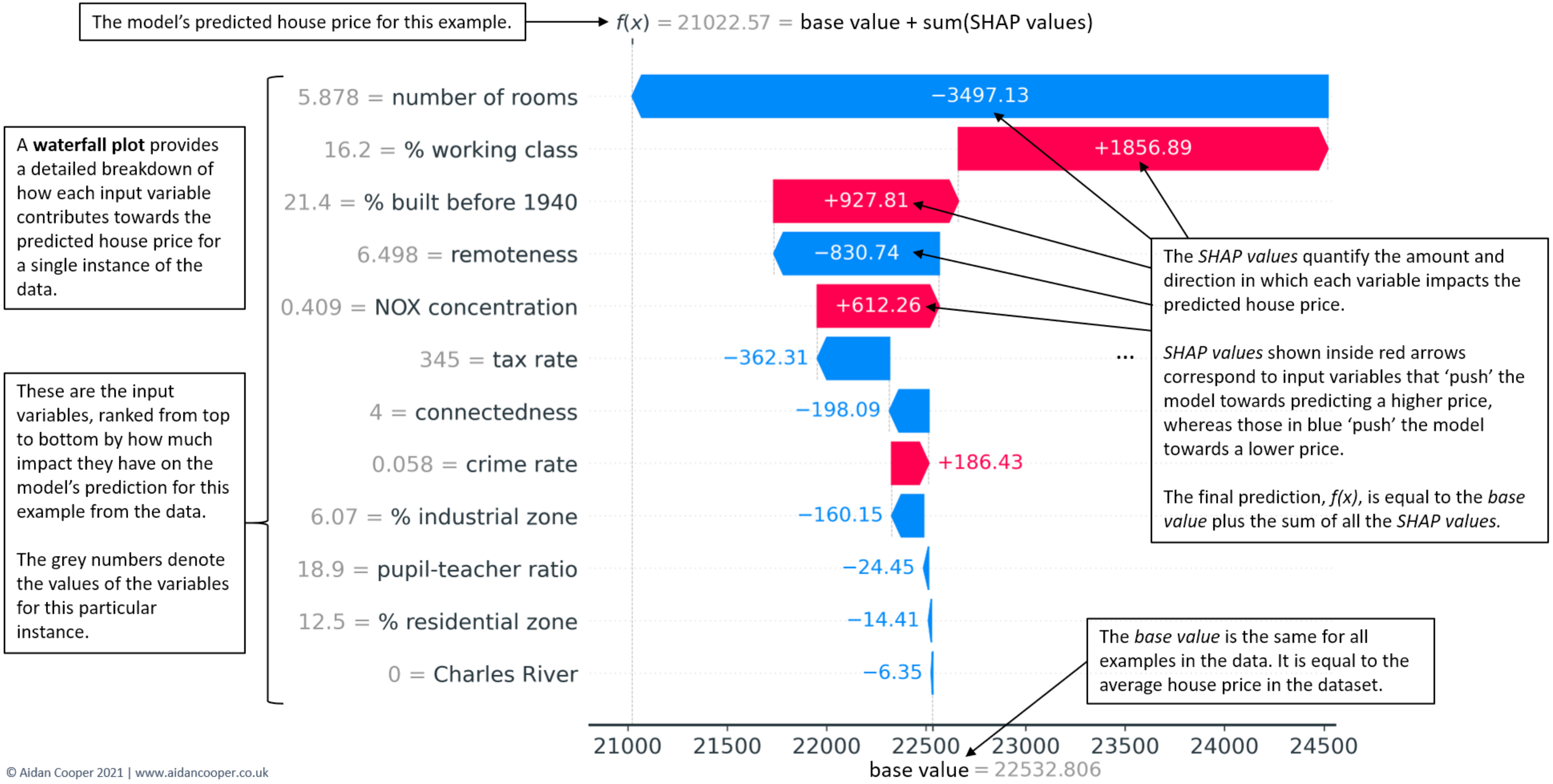
A machine learning model's prediction, *f(x)*, can be represented as the sum of its computed *SHAP values*, plus a fixed *base value*, such that:

$$f(x) = base\ value + sum(SHAP\ values)$$


Regression:
var1 var2 var3 ...
House1
House2
House3
...
DATA
ML Regressor
f(x) = base value + sum(SHAP values)
$25,000 = $22,500 + sum(House1 SHAP values)
$36,500 = $22,500 + sum(House2 SHAP values)
$29,000 = $22,500 + sum(House3 SHAP values)
SHAP
var1 var2 var3 ...
House1
House2
House3
...
SHAP Values



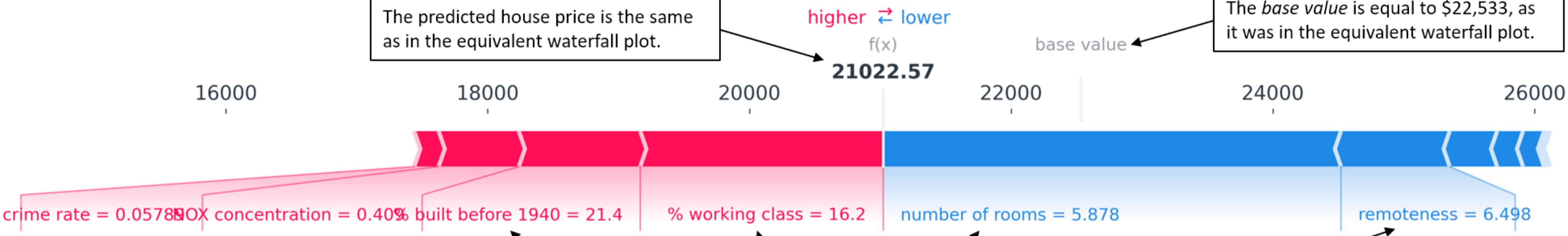
Model Input Variables
Target Variable
count of instances
crime rate
% residential zone
% industrial zone
Charles River
NOX concentration
number of rooms
% built before 1940
remoteness
connectedness
tax rate
pupil-teacher ratio
% working class
predicts...
house price ($)
50,000
40,000
30,000
20,000
10,000


The model’s predicted house price for this example.
f(x) = 21022.57 = base value + sum(SHAP values)
A **waterfall plot** provides a detailed breakdown of how each input variable contributes towards the predicted house price for a single instance of the data.
These are the input variables, ranked from top to bottom by how much impact they have on the model’s prediction for this example from the data.
The grey numbers denote the values of the variables for this particular instance.
5.878 = number of rooms
−3497.13
16.2 = % working class
+1856.89
21.4 = % built before 1940
+927.81
6.498 = remoteness
−830.74
0.409 = NOX concentration
+612.26
345 = tax rate
−362.31
4 = connectedness
−198.09
0.058 = crime rate
+186.43
6.07 = % industrial zone
−160.15
18.9 = pupil-teacher ratio
−24.45
12.5 = % residential zone
−14.41
0 = Charles River
−6.35
...
The *SHAP values* quantify the amount and direction in which each variable impacts the predicted house price.
SHAP values shown inside red arrows correspond to input variables that ‘push’ the model towards predicting a higher price, whereas those in blue ‘push’ the model towards a lower price.
The final prediction, *f(x)*, is equal to the *base value* plus the sum of all the *SHAP values*.
The *base value* is the same for all examples in the data. It is equal to the average house price in the dataset.
21000
21500
22000
22500
23000
23500
24000
24500
base value = 22532.806
© Aidan Cooper 2021 | www.aidancooper.co.uk

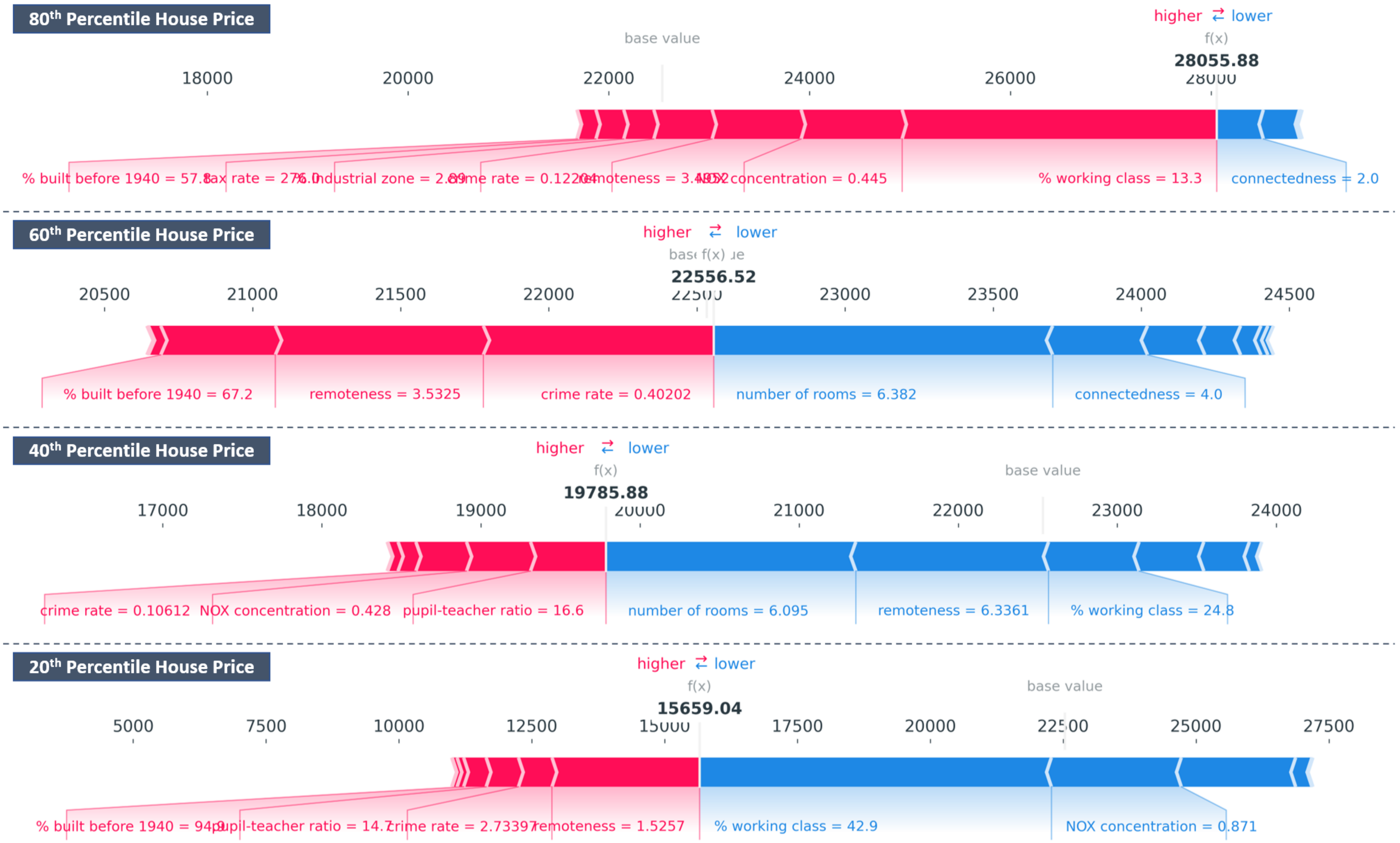
Force Plot

The predicted house price is the same as in the equivalent waterfall plot.

The *base value* is equal to $22,533, as it was in the equivalent waterfall plot.



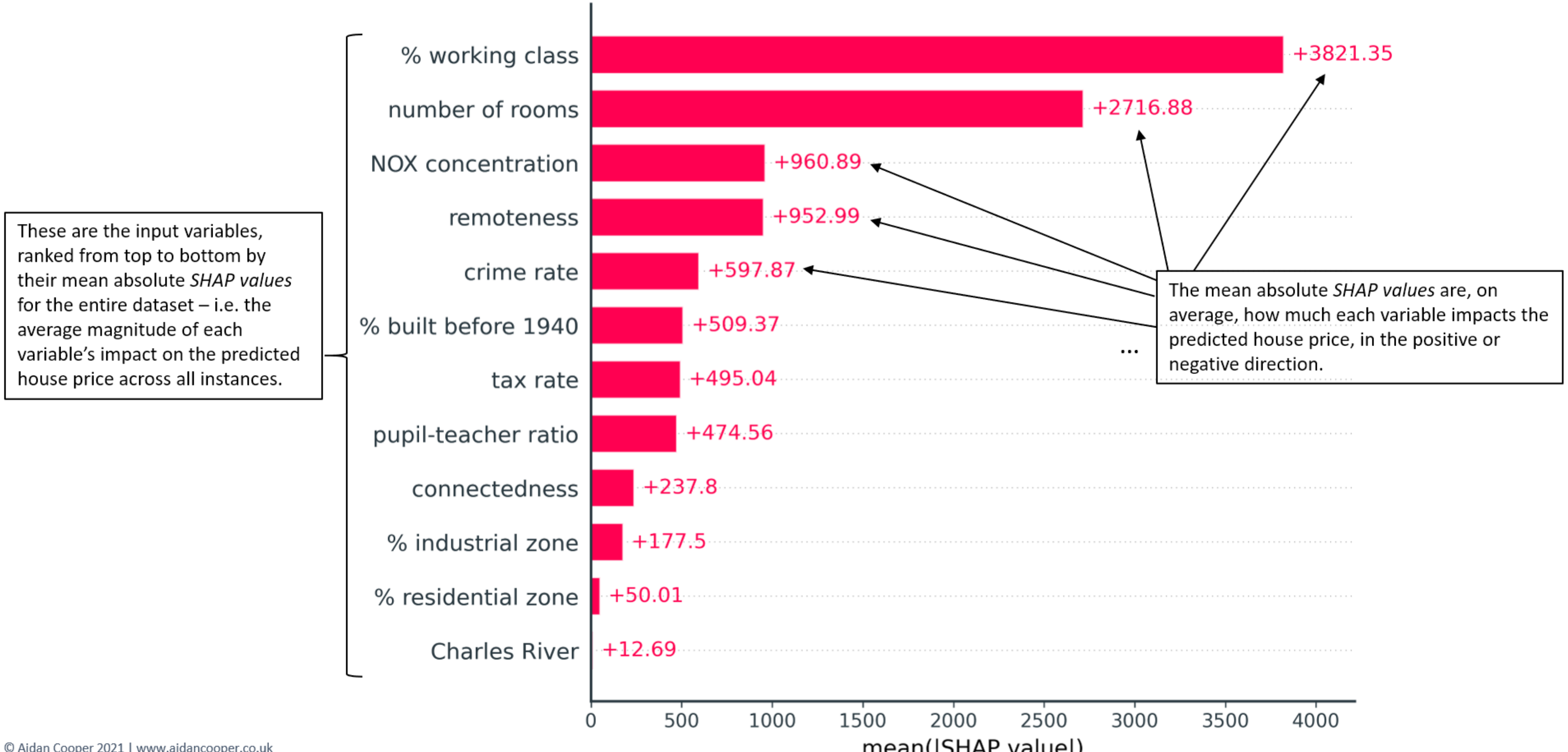
In a **force plot**, variables with *SHAP values* that 'push' the model towards a higher price appear on the left in red, whereas those that 'push' the model towards a lower price appear on the right in blue. The actual value of the variable is shown alongside the variable name. Variables with larger *SHAP values* (i.e. more impact) have larger arrows.



80th Percentile House Price
higher ⇄ lower
base value
f(x)
28055.88
18000
20000
22000
24000
26000
28000
% built before 1940 = 57.8
tax rate = 276.0
% industrial zone = 2.89
crime rate = 0.12204
remoteness = 3.4952
NOX concentration = 0.445
% working class = 13.3
connectedness = 2.0
60th Percentile House Price
higher ⇄ lower
base value
f(x)
22556.52
20500
21000
21500
22000
22500
23000
23500
24000
24500
% built before 1940 = 67.2
remoteness = 3.5325
crime rate = 0.40202
number of rooms = 6.382
connectedness = 4.0
40th Percentile House Price
higher ⇄ lower
f(x)
19785.88
base value
17000
18000
19000
20000
21000
22000
23000
24000
crime rate = 0.10612
NOX concentration = 0.428
pupil-teacher ratio = 16.6
number of rooms = 6.095
remoteness = 6.3361
% working class = 24.8
20th Percentile House Price
higher ⇄ lower
f(x)
15659.04
base value
5000
7500
10000
12500
15000
17500
20000
22500
25000
27500
% built before 1940 = 94.9
pupil-teacher ratio = 14.7
crime rate = 2.73397
remoteness = 1.5257
% working class = 42.9
NOX concentration = 0.871

Global interpretability: understanding drivers of predictions across the population

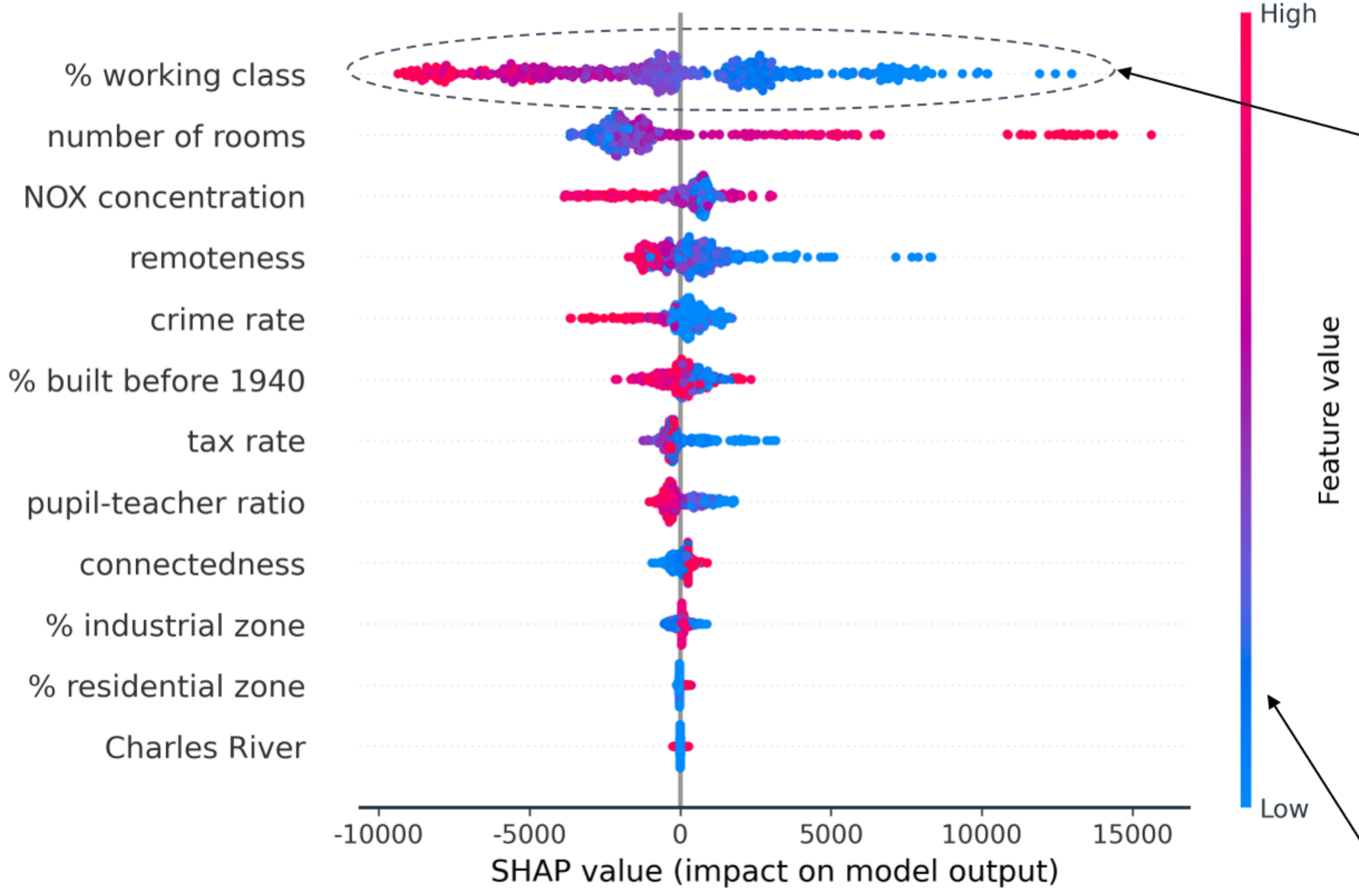
- Bar Plot



Beeswarm plots

These are the input variables, ranked from top to bottom by their mean absolute *SHAP values* for the entire dataset.

Note: this ranking is exactly the same as for the bar plot.



In a **beeswarm plot**, for each variable, every instance (i.e. row) of the dataset appears as it's own point. The points are distributed horizontally along the *x*-axis according to their *SHAP* value. In places where there is a high density of *SHAP values*, the points are stacked vertically.

Examining how the *SHAP* values are distributed reveals how a variable may influence the model's predictions.

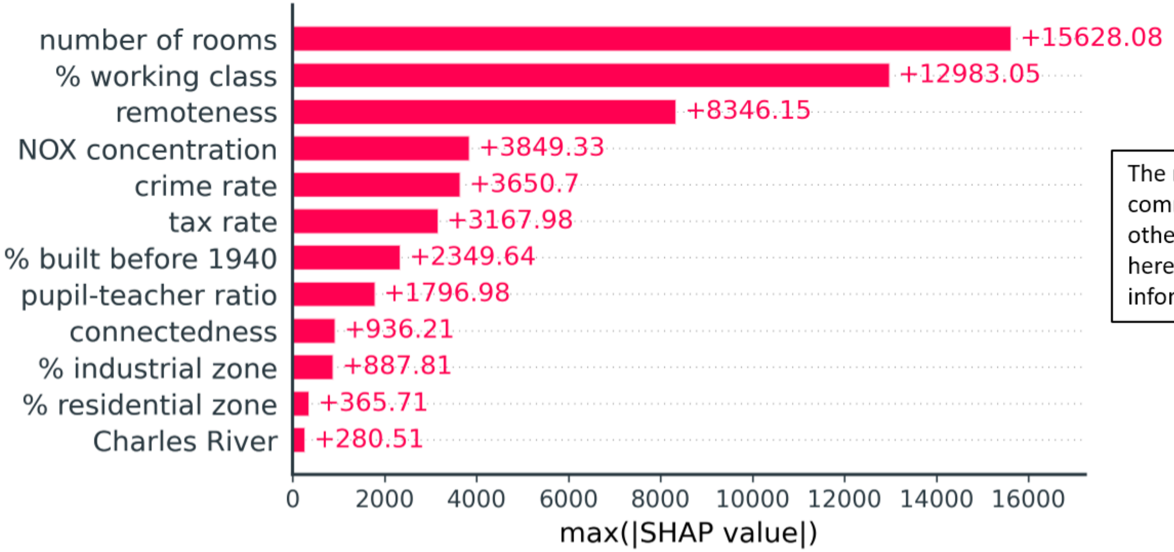
The colour bar corresponds to the raw values (not to be confused with the *SHAP values*) of the variables for each instance (i.e. point) on the graph.

If the value of a variable for a particular instance is relatively high, it appears as a red dot. Relatively low variable values appear as blue dots.

Examining the colour distribution horizontally along the *x*-axis for each variable provides insights into the general relationship between a variable's raw values and its *SHAP values*.

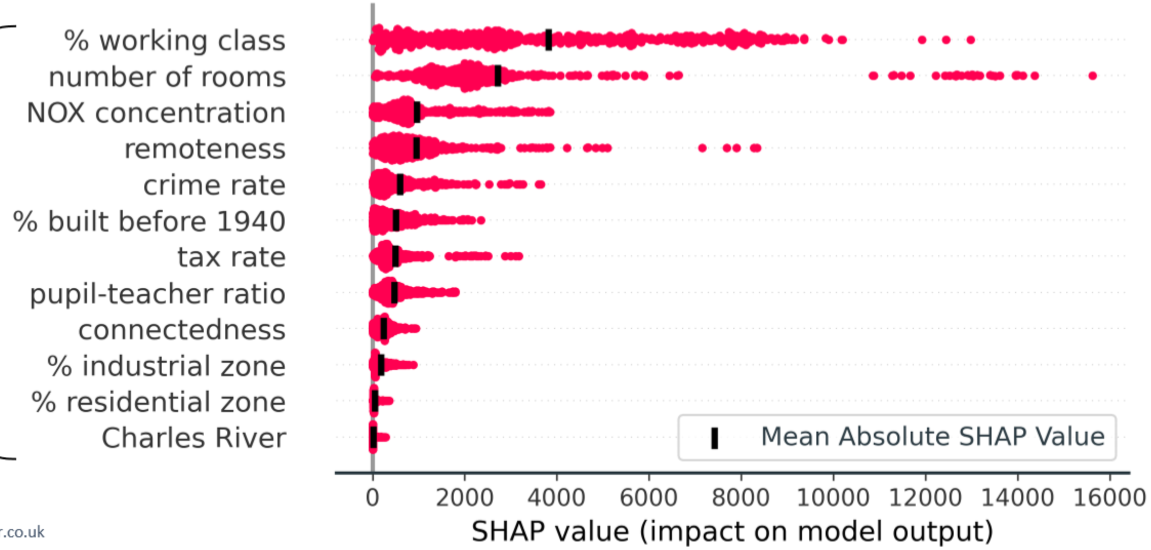
Whereas before, variables were ranked by their **mean** absolute *SHAP value*, here they are ranked by their **max** absolute *SHAP value* for the entire dataset.

Note that the ranking changes in places.



The mean absolute *SHAP value* is the most commonly used ranking for variables, but other statistics such as the max value (shown here) or median value may also be informative.

These are the input variables, ranked from top to bottom by their mean absolute *SHAP values* for the entire dataset.



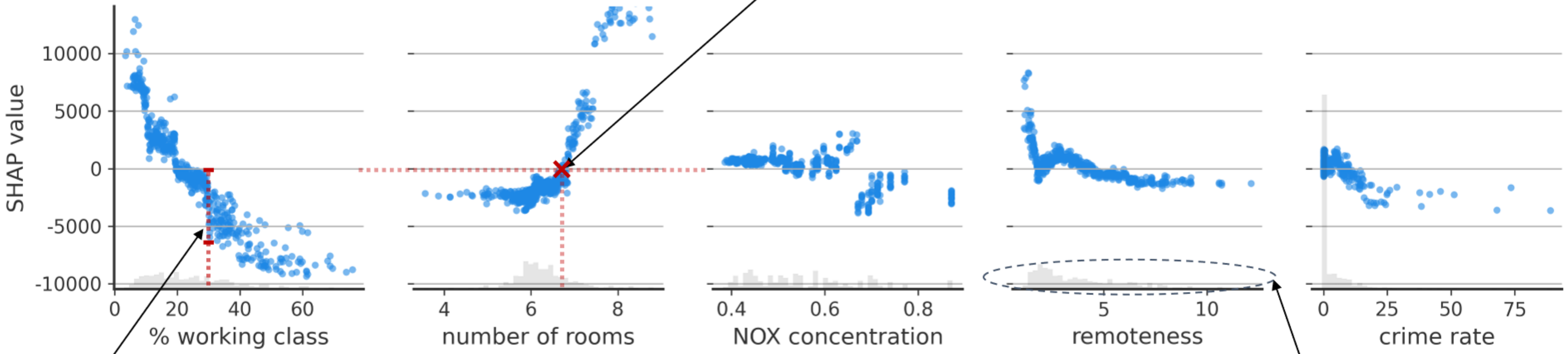
A balance can be struck between the simplicity of a bar plot and the information-rich complexity of a beeswarm plot, by creating a beeswarm plot for the **absolute** *SHAP values*.

This still shows us the ranking and relative influence of variables on the model's predictions, but also allows for further insights. E.g. the highest observed *SHAP values* actually occur for the 2nd ranked variable, *number of rooms*.

In a **dependence plot**, every instance (i.e. row) of the dataset appears as it's own point. The points are presented as a scatterplot of a variable's *SHAP values* versus the variables underlying raw values.

SHAP values above the *y*=0 line lead to predictions of higher house prices, whereas those below it are associated with lower house price predictions. The raw variable value at which the distribution of *SHAP values* cross the *y*=0 line tells you the threshold at which the model switches from predicting lower to higher house prices. For *number of rooms*, this is at approximately 6.8 rooms, as marked by the ✕.

With all five plots on the same *y*-scale, the extent of the vertical distribution of the *SHAP values* indicates how much relative influence each variable has on predictions. *% working class* has a much wider range of *SHAP values* than *crime rate*.



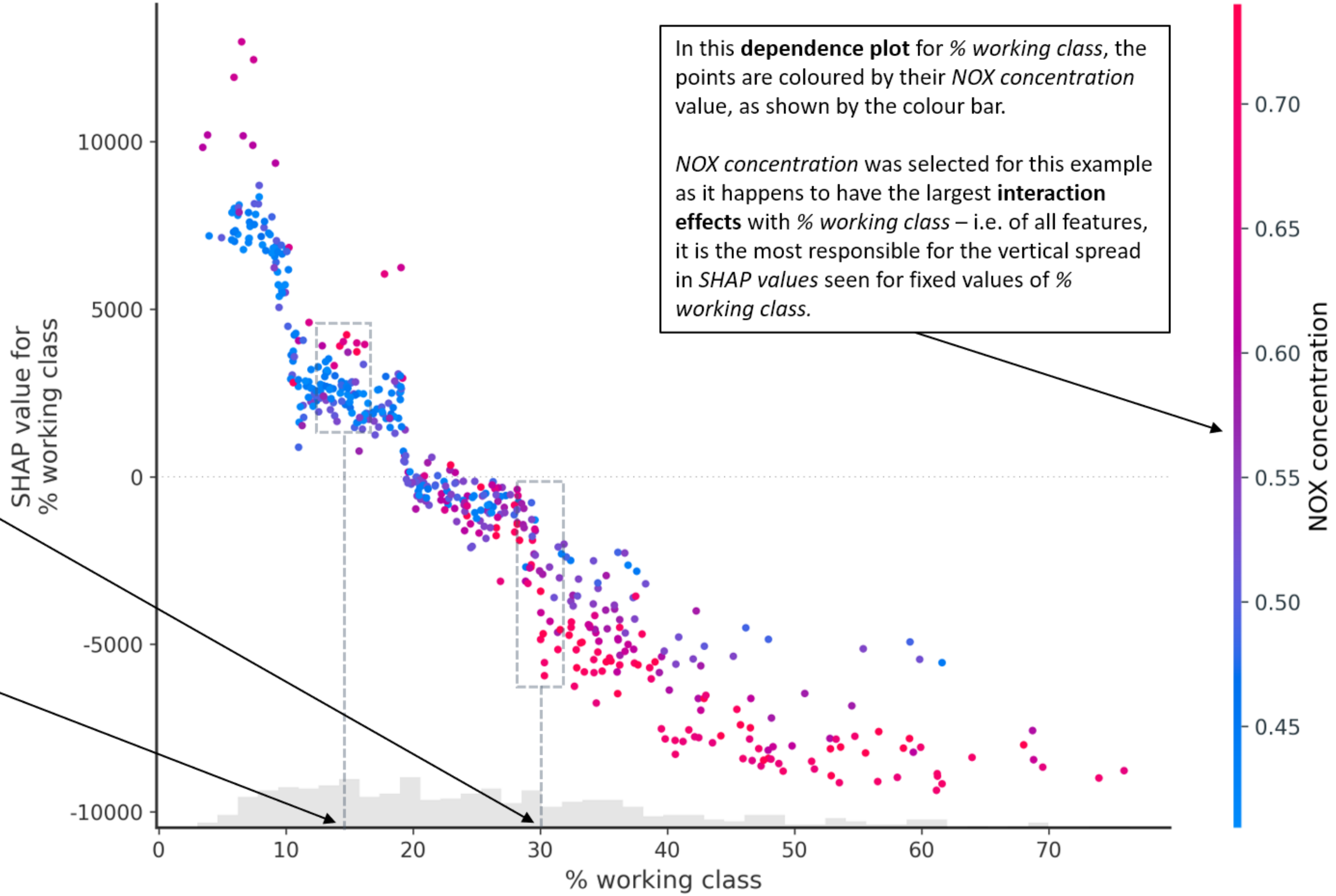
The vertical spread of *SHAP values* at a fixed raw variable value is due to *interaction effects* with other variables. For example, here we see that houses with a *% working class* of 30% can have *SHAP values* that range from $0 to -$6,500 depending on the other data for those particular instances.

The shapes of the distributions of points provide insights into the relationship between a variable's values and its *SHAP values.* For *% working class*, we see a negative, linear relationship across the full range of variable values. For *number of rooms*, we see that *SHAP values* are mostly flat between 4 and 6.5 rooms, but then increase sharply for higher room counts.

The inset histograms just above the *x*-axis display the distributions of raw variable values. We should be cautious not to overinterpret regions of the dependence plot where the underlying data is sparse (e.g. *crime rates* over 25%).

The *SHAP values* at a *% working class* value of 30% range from roughly $0 to -$6,500. Examining the colour distribution, it can be seen that higher levels of *NOX concentration* are associated with the decrease in *SHAP values* for this region.

At a *% working class* value of 15%, the *SHAP values* range from roughly $1,000 to $5,000. For this region, it can be seen that instances with the highest *NOX concentration* values have the highest *SHAP values*, whereas instances with lower *NOX concentration values* have lower *SHAP values*. This is the inverse relationship to that seen at 30% working class.



In this **dependence plot** for *% working class*, the points are coloured by their *NOX concentration* value, as shown by the colour bar.

NOX concentration was selected for this example as it happens to have the largest **interaction effects** with *% working class* – i.e. of all features, it is the most responsible for the vertical spread in *SHAP values* seen for fixed values of *% working class.*