EXPLICABILITY/XAI

INTUITION

- Principle: Using supervised machine learning to better understand a dataset

- Correlation: if X increases, Y increases
  - Supervised ML: How X can be used to predict Y
    - =>Extract from the model the relation between X and Y
    - =>Take into account also *interactions* with other variables

INTUITION

- Two aspects of interpretability
  - 1) Feature Importance
    - How much X impacts Y?
  - 2)Nature of the relation
    - When X increases, how does Y change?
    - In general?
    - For a particular observation?
    - Depending on another variable?

- Example
  - The price of an apartment may depend on the floor
    - Negative relation for low values and positive for high values
    - Depends on the value of the variable "Elevator"
    - For a particular appartment (beutiful view…), a particular relation.

AD-HOC

- Interpretable methods
  - Linear/Logistic regressions
    - Lasso
  - Decision tree (small)
  - K-NN
  - …

- Black boxes
  - Random forests/XGBoost, etc.
  - Deep Neural Network
  - …

AD-HOC

- Example: Linear regression
  - The coefficients == explain relation between variables and target
    - More powerful than correlation coefficient (take other variables into account)
    - Smoking causes cancer. Old people smoke less. Old people have more cancer
    - =>Simple correlation: smoking has little effect on cancer, or even negative correlation
    - =>Linear regression parameters: smoking increases cancer, age increases cancer
  - Variable importance/impact ?
  - /!\ Be careful to raw values!
    - 1 cigarette increases cancer rate by…
    - 1 year increases cancer rate by…
    - Not comparable
    - =>Normalize the variables
    - Then you can compare the coefficient values

AD-HOC

- Decision tree/Regression Tree
  - Relation variable/Target:
    - Can be read in the tree
    - =>If the building has an elevator, then… else…
  - Feature importance
    - Computed from the gain in the objective
    - => sklearn: tree.feature_importances

AD-HOC

- Computing feature importance in a tree
  - $\text{FI}(f) = \sum_{n \in N(f)} \frac{N_n}{N} \Delta I_n$
    - $N(f)$: set of internal nodes splitting on feature f
    - N_n number of training samples reaching node n
    - N total number of samples
    - ΔI_n Objective decrease produced by that split (RMSE, Gini, etc.)

- Can be normalized to sum to 1

AGNOSTIC FEATURE IMPORTANCE

Permutation Feature Importance

AGNOSTIC METHODS

- Evaluating feature importance even in black-box models
  - Independent of the ML algorithm
- Global score, all cases together

PERMUTATION FEATURE IMPORTANCE

- Intuition: If a variable is important for a model, removing it reduces the performance of the model
  - Feature importance == how much performance is lost without this feature?

- How to remove the feature without changing the model?
  - Randomize the values of the feature

- Score: $FI_j = s(f(D)) - s(f(D\pi_j))$
  - s: scoring function
  - f: ML model
  - D: dataset
  - $D\pi_j$: Dataset with variable j randomized

XAI

MOTIVATION

- *Statistical Modeling: The Two Cultures*
  - L. Breiman (2001), Statistical Science,
- Historically, in data analysis, two cultures
  - Model-based: we assume data follows some statistical model
    - Interpretable methods,
    - Limited complexity
    - A priori on the data (human intuition)
  - Algorithmic/ML
    - Focus on prediction accuracy
- XAI: uniting the two

XAI

- Field concerned with making outcomes yielded by black box models interpretable

- Motivations:
  - Naturally interpretable models are usually more “naive”, have lower capacity of expression
    - =>We want to keep the full power of black-box methods, while being able to explain decisions
    - => e.g.: European Union directive: AI models used to take decision must be able to explain that decision
  - To understand relations between variables:
    - If the relation is really complex, the simplified version by a more naive method will be les accurate.

LIME

- LIME = Local Interpretable Model-agnostic Explanations
  - Ribeiro, Singh & Guestrin, 2016, KDD
- Idea: approximate the black box locally by a simple model
  - Explain the decision for instance x_0
    - For one apartment x_0, the model answered Y.
    - =>What was the role of each variable in this decision?
    - =>For instance: for this apartment, the floor played a positive role…

LIME

- Principle: Builds a **surrogate model**, valid locally
  - Surrogate model: A simpler model (linear regression, decision tree) that mimic the behavior of the complex model
    - Fitted to predictions of the model, not to real data
  - Intuition: In the solution space, we need a complex model (elevator/no elevator, each city, old/new buildings…)
    - => But locally (e.g., Haussmanian bulding in Paris with no elevator), the model can be well approximated by a simpler one

LIME

- Local model behavior approximation
  - Generate random, synthetic points
    - Random perturbations of the point of interest
  - Approximate with a simple model
    - e.g., linear regression
  - Use a loss weighted for proximity
    - More similar points count more

LIME

- $\text{argmin}_g \; L(f, g, \pi x_0) + \Omega(g)$
  - g: local surrogate model
  - f: model to approximate
  - L: local loss
  - πx_0: Locality kernel: control the similarity of sampled points
  - $\Omega(g)$: complexity penalty (keep the model simple, regularization)

LIME

- Surrogate model:
  - A linear or logistic regression model
  - A tree of small size
- Regularized, custom loss for weighting more the less perturbated points
- The model can be interpreted as usual

SHAP

- SHapley Additive exPlanations
  - Lundberg & Lee, 2017, NeurIPS: "A Unified Approach to Interpreting Model Predictions"

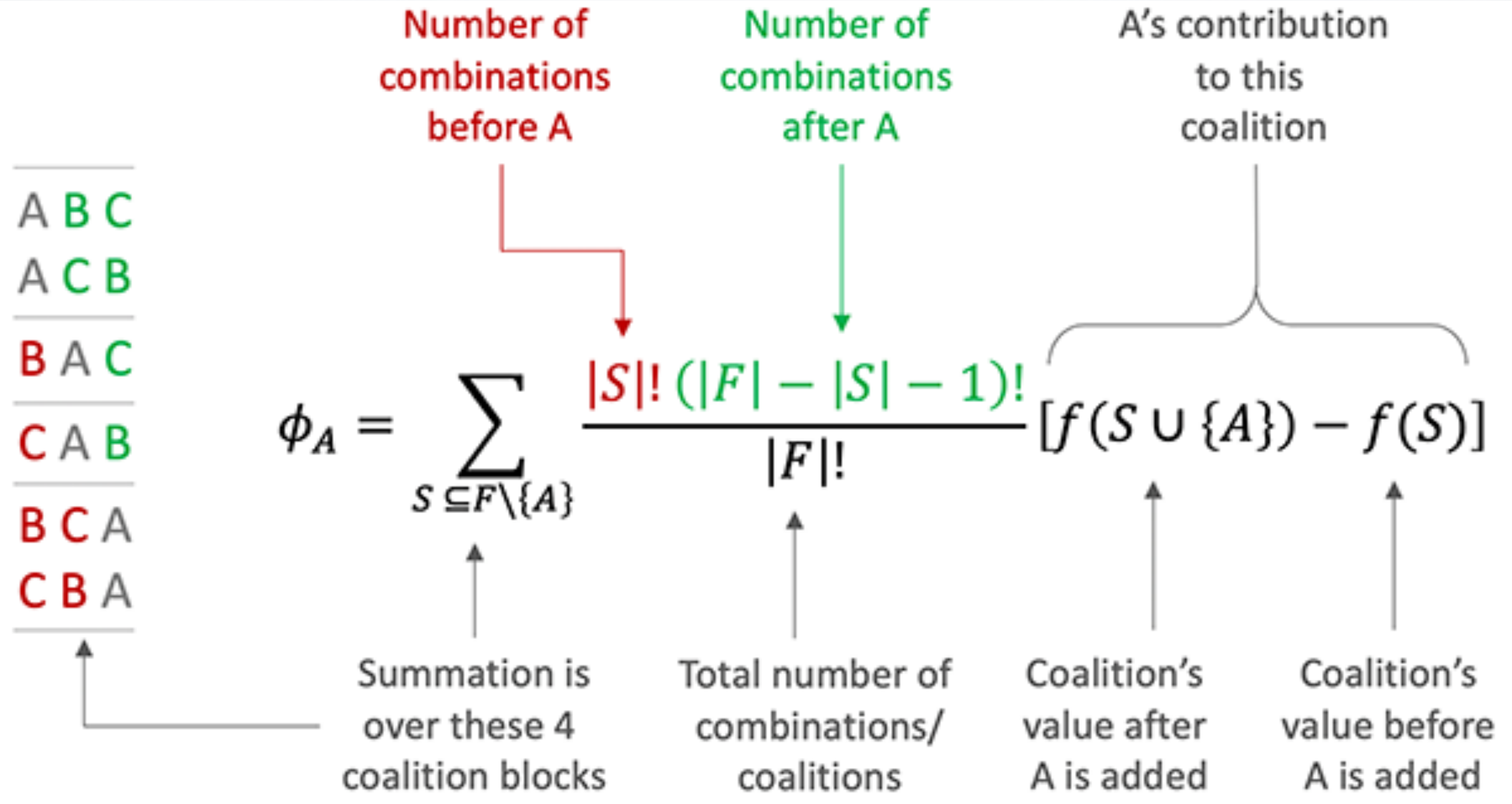
- Principle:
  - Local estimate (for one prediction)
  - Observe how much each feature **changes the outcome** when it is added to **a subset of other variables**
    - Adding *age* to predict *cancer* change has different effect if we already include *smoking* or not

SHAP

- $$\phi_i(x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$
  - F: set of all features
  - $S \subseteq F \setminus \{i\}$: subset of features without feature of interest i ("coalition")
  - $f_S(x_S)$: model output when using only features in S
  - $f_{S \cup \{i\}}(x_{S \cup \{i\}})$: model outcome when using featues in S and i
  - $\phi_i(x)$: SHAP value for variable i for observation x
- The large term with factorials is just a weighting to account for multiple possible combinations leading to the same case

SHAP

Compute effect of variable A



https://www.linkedin.com/pulse/shap-bridging-gap-between-machine-predictions-igor-alcantara/

SHAP

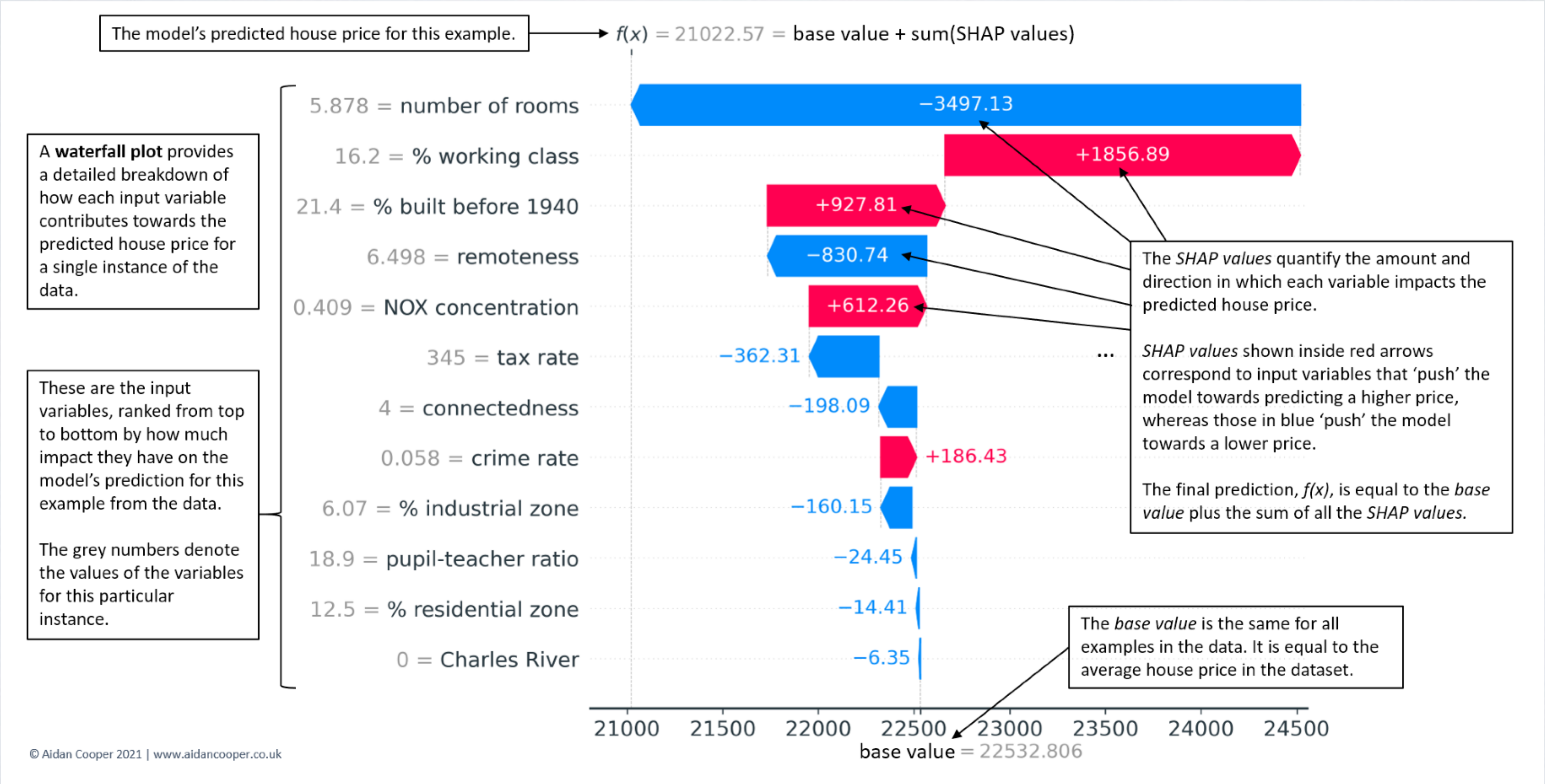
- In practice:
  - Removing variables means replacing original values with random values (e.g., taking values at random in the dataset for this variable)
  - High complexity: impossible to compute in full in practice
    - Approximate by sampling

SHAP

- Four axioms of SHAP
  - **Efficiency:** Total of all feature contributions equals the actual model output
  - **Symmetry:** If two features contribute identically, they get identical SHAP values
  - **Dummy:** If a feature never changes the outcome, then it gets SHAP value $\phi_i = 0$
  - **Additivity:** if two models f, g are combined linearly $(h = f + g)$, then $\phi_i(h) = \phi_i(f) + \phi_i(g)$

SHAP: GLOBAL IMPORTANCE

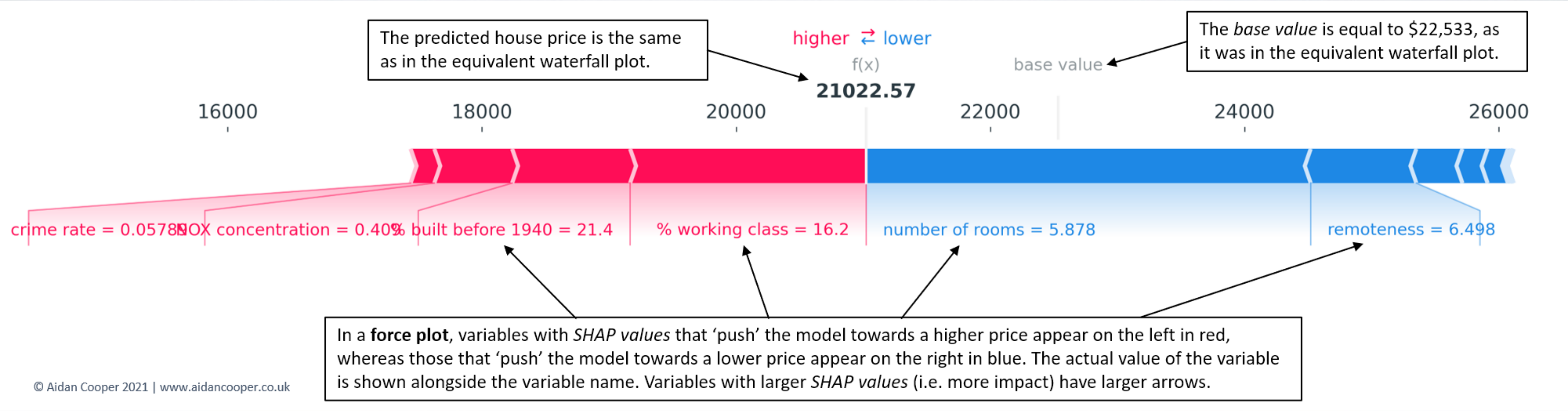
Individual instance explanation



https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/

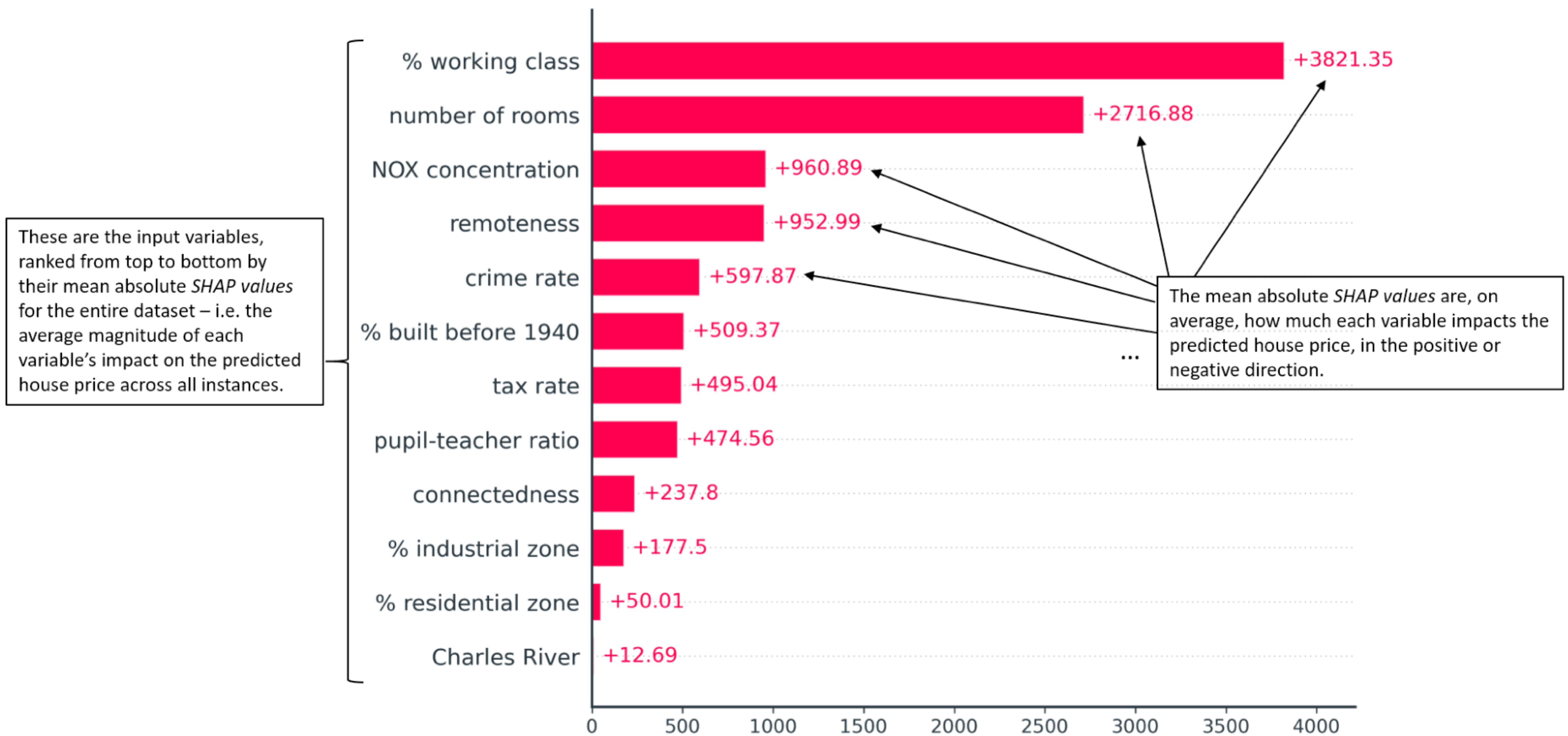
SHAP: GLOBAL IMPORTANCE

Individual instance explanation



SHAP: GLOBAL IMPORTANCE

Global feature importance



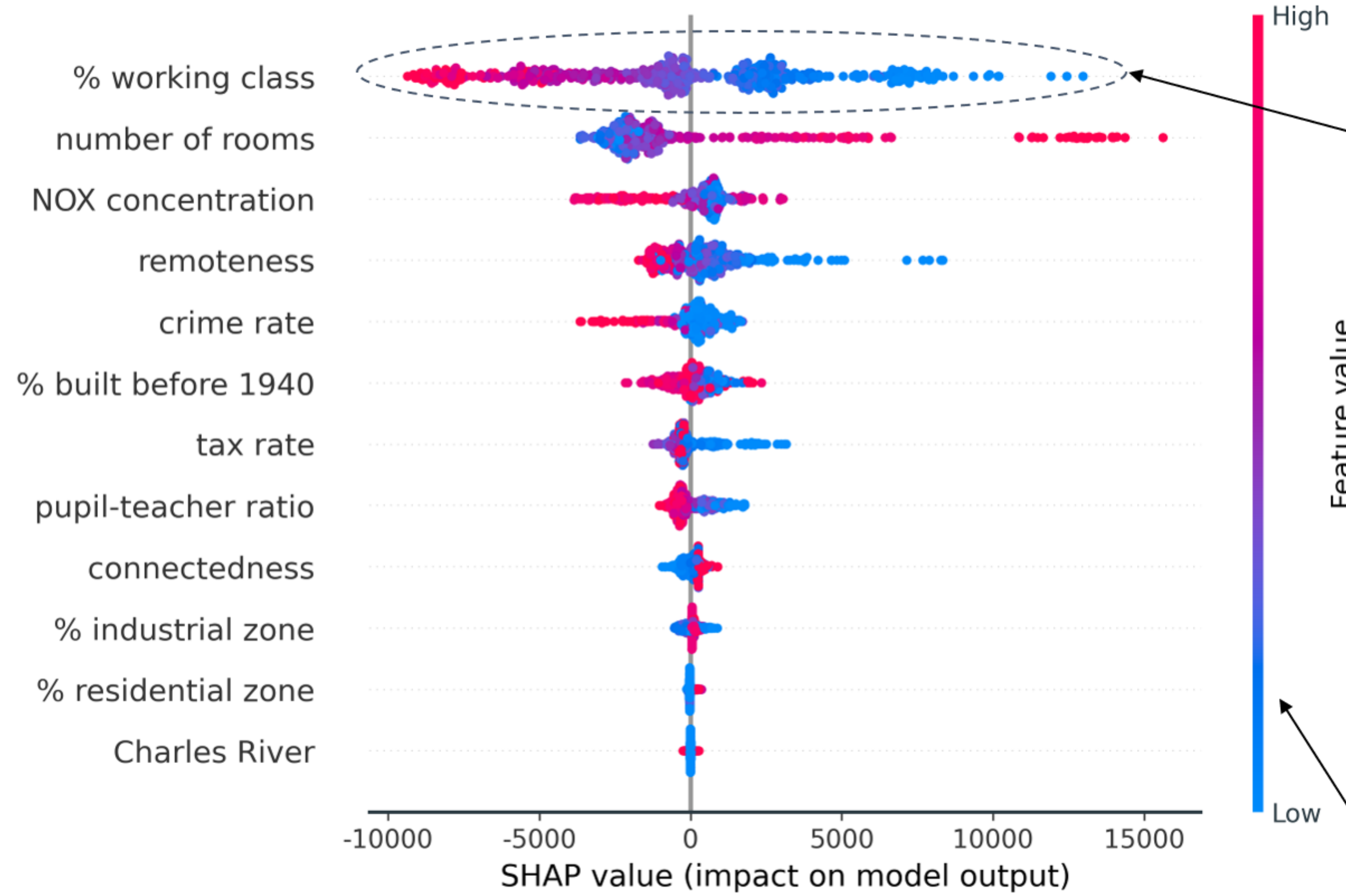
© Aidan Cooper 2021 | www.aidancooper.co.uk

SHAP: GLOBAL IMPORTANCE

Beeswarm plot

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*.

© Aidan Cooper 2021 | www.aidancooper.co.uk

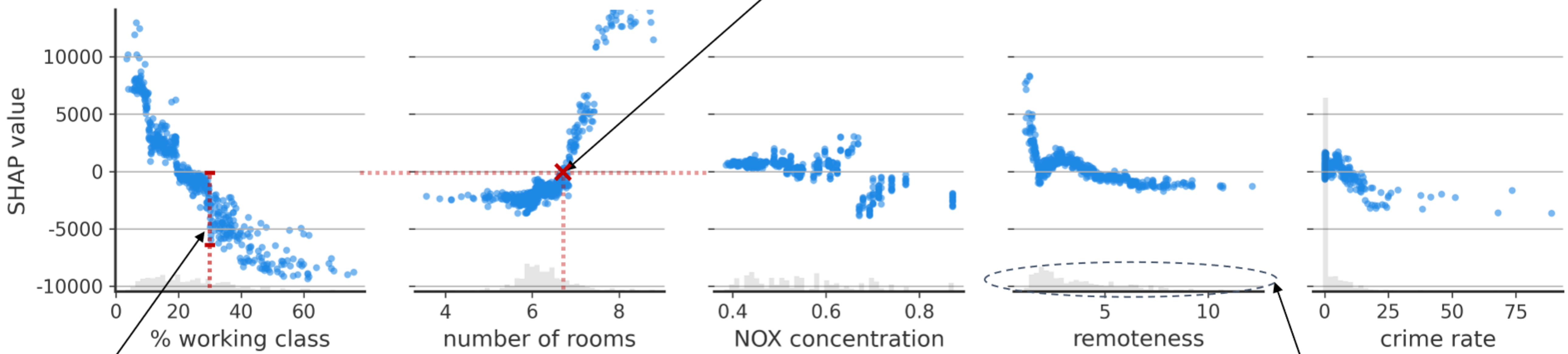
SHAP: GLOBAL IMPORTANCE

Dependence plot

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.

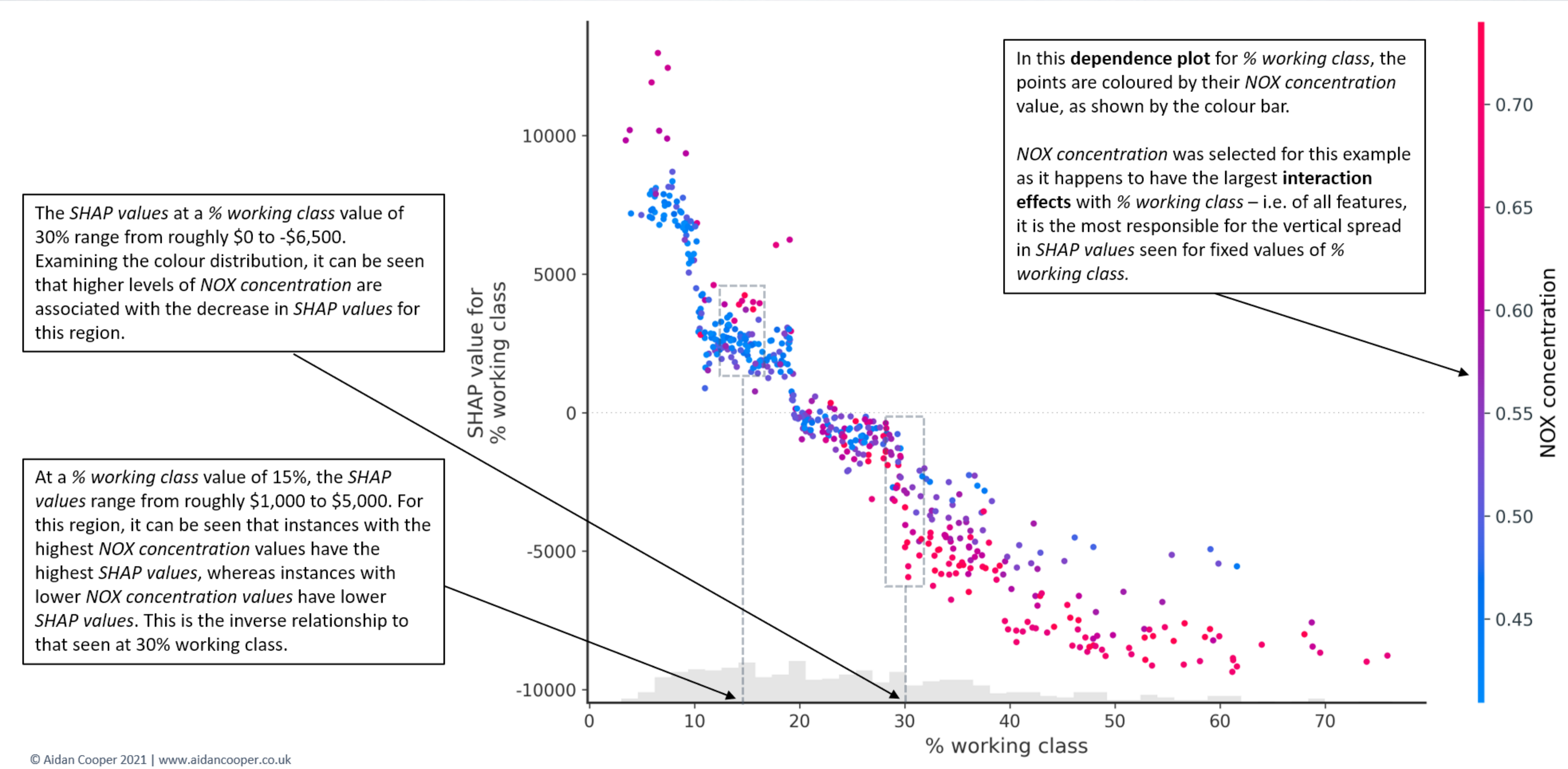
The shapes of the distributions of points provide insights into the relationship between a variable's values and its *SHAP values.* For *% working class*, we see a negative, linear relationship across the full range of variable values. For *number of rooms*, we see that *SHAP values* are mostly flat 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%).

© Aidan Cooper 2021 | www.aidancooper.co.uk

SHAP: GLOBAL IMPORTANCE

Interaction plot



© Aidan Cooper 2021 | www.aidancooper.co.uk