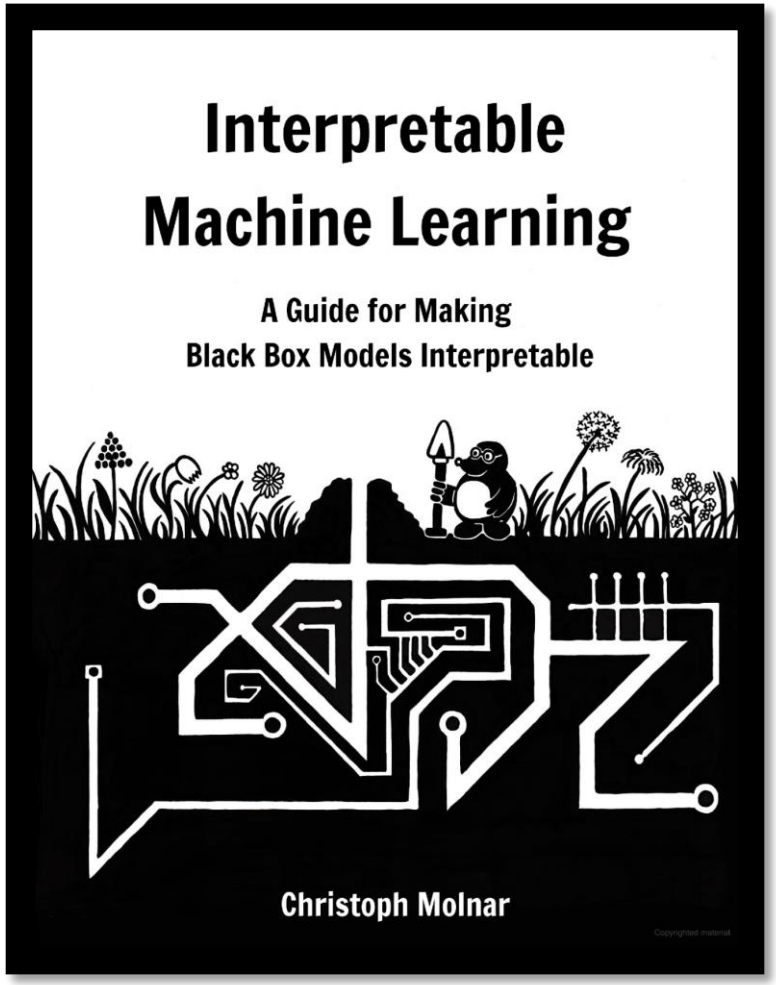
Interpretable Machine Learning

Reference Book

- Molnar, Christoph. *Interpretable machine learning*. Lulu. com, 2020.



Interpretability

- It is difficult to (mathematically) define interpretability.

- A (non-mathematical) definitions:
  - **Interpretability is the degree to which a human can understand the cause of a decision.**
  - **Interpretability is the degree to which a human can consistently predict the model's result.**

Explanation Models

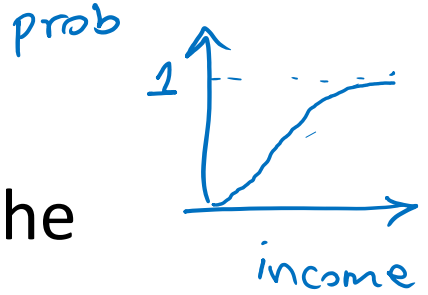
- Global vs. Local

- Model-specific vs. Model-agnostic

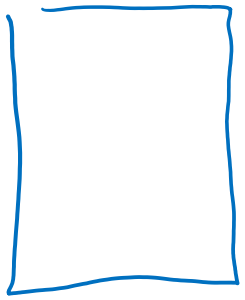
Global Model-Agnostic Methods

- **partial dependence plot**: feature effect method
- **Accumulated local effect plots**
- **Feature interaction (H-statistic)**
- **Functional decomposition**
- **Permutation feature importance**: measures the importance of a feature as an increase in loss when the feature is permuted.
- **Global surrogate models**: replaces the original model with a simpler model for interpretation
- **Prototypes and criticisms**

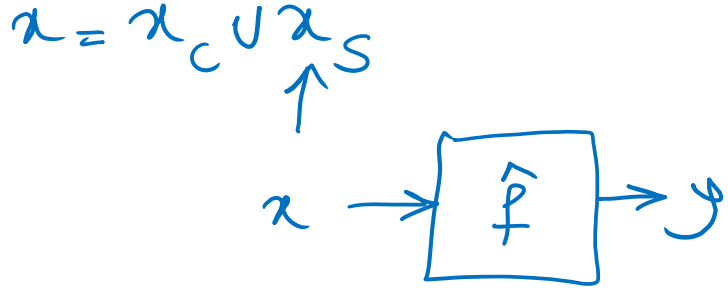
Partial Dependence Plot (PDP)



- PDP shows the marginal effect one or two features have on the predicted outcome of a machine learning model.



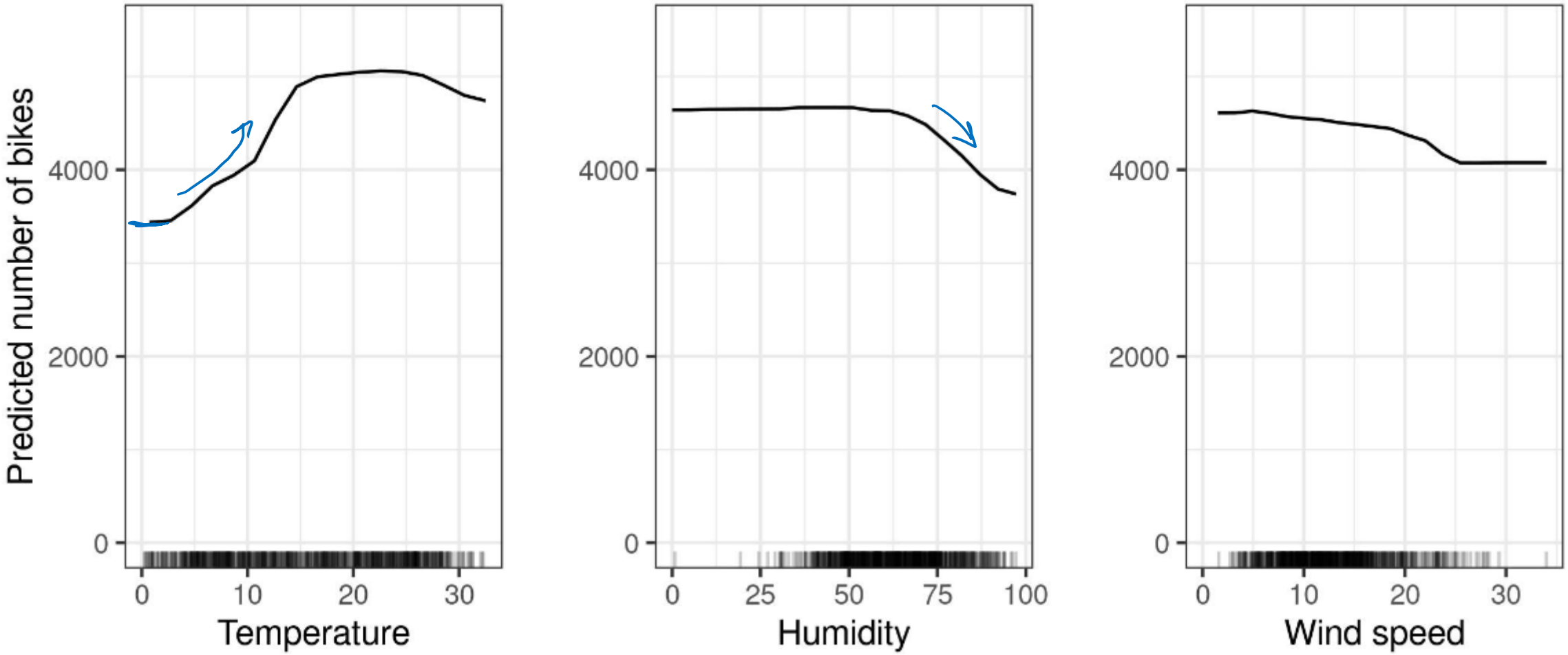
$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_S, x_C^{(i)})$$



- The partial function tells us for given value(s) of features S what the average marginal effect on the prediction is.
- In this formula, $x_C^{(i)}$ are actual feature values from the dataset for the features in which we are not interested, and n is the number of instances in the dataset.

PDP Example

bike



Permutation feature importance

Input: Trained model $\hat{f}$, feature matrix X, target vector y, error measure $L(y, \hat{f})$.

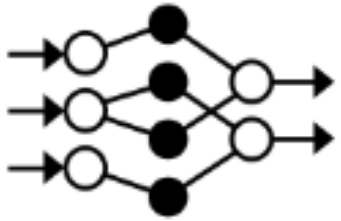
1. Estimate the original model error $e_{orig} = L(y, \hat{f}(X))$ (e.g. mean squared error)
2. For each feature $j \in \{1, \ldots, p\}$ do:
   - Generate feature matrix X_{perm} by permuting feature j in the data X. This breaks the association between feature j and true outcome y.
   - Estimate error $e_{perm} = L(Y, \hat{f}(X_{perm}))$ based on the predictions of the permuted data.
   - Calculate permutation feature importance as quotient $FI_j = e_{perm}/e_{orig}$ or difference $FI_j = e_{perm} - e_{orig}$
3. Sort features by descending FI.

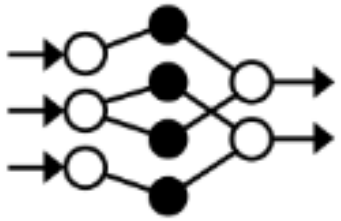
Permutation feature importance

	Feature_1	Feature_2	Feature_3	Label
Sample_1	1970	10.5	1	403.12
Sample_2	2020	14.9	2	412.15
Sample_3	1910	17.7	3	564.46

Permute "*Feature_2*"

	Feature_1	Feature_2	Feature_3	Label
Sample_1	1970	17.7	1	403.12
Sample_2	2020	10.5	2	412.15
Sample_3	1910	14.9	3	564.46





Price (Predicted Original)
350.54
420.24
600.13

Price (Predicted Permuted)
700.65
300.13
121.15

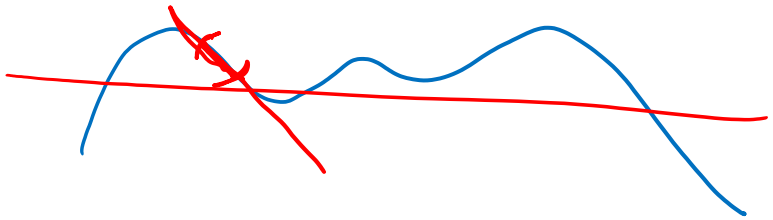
Average difference
= Feature Importance

Global surrogate models

- Provide an interpretable model interpreting the original complex model.

1. Input data X into complex model (e.g. neural net)
2. Generate predictions - Y_{pred}
3. Train interpretable model on X and Y_{pred}

Global Surrogate Model



$x^1, x^2, \cdots, x^N$ → Black Box → $y^1, y^2, \cdots, y^N$

(e.g. Neural Network)

$x^1, x^2, \cdots, x^N$ → Linear Model → $\tilde{y}^1, \tilde{y}^2, \cdots, \tilde{y}^N$

$\cdots\cdots$

as close as possible

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

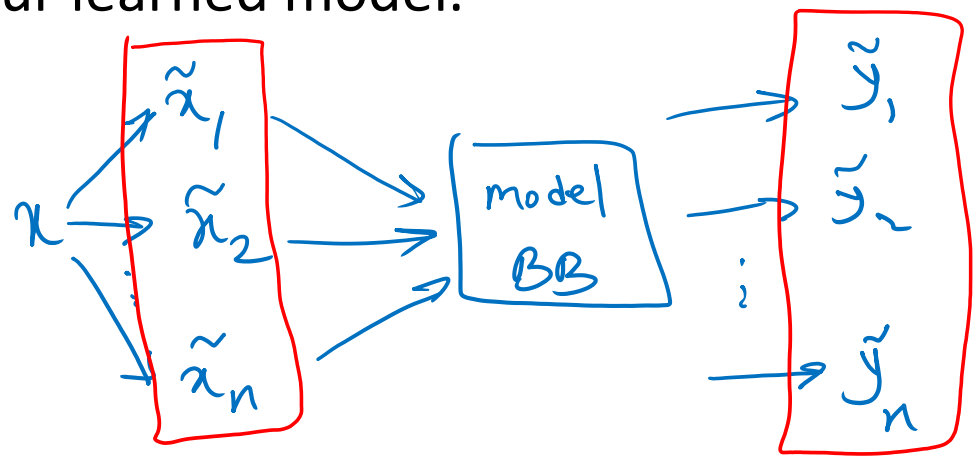
Local Model-Agnostic Methods

- Individual conditional expectation
- Local surrogate models (LIME)
- Scoped rules (anchors)
- Counterfactual explanation
- Shapley values
- SHAP is another computation method for Shapley values, but also proposes global interpretation methods based on combinations of Shapley values across the data.

LIME

- Local Surrogate: LIME (Local Interpretable Model-agnostic Explanation)

- Explain a single (local) observation/example:
  1. Select single observation
  2. Use original black-box model to generate predictions near the observation by perturbing it.
  3. Fit a linear (or an interpretable) model.
  4. Interpret your learned model.



Intuition Behind LIME

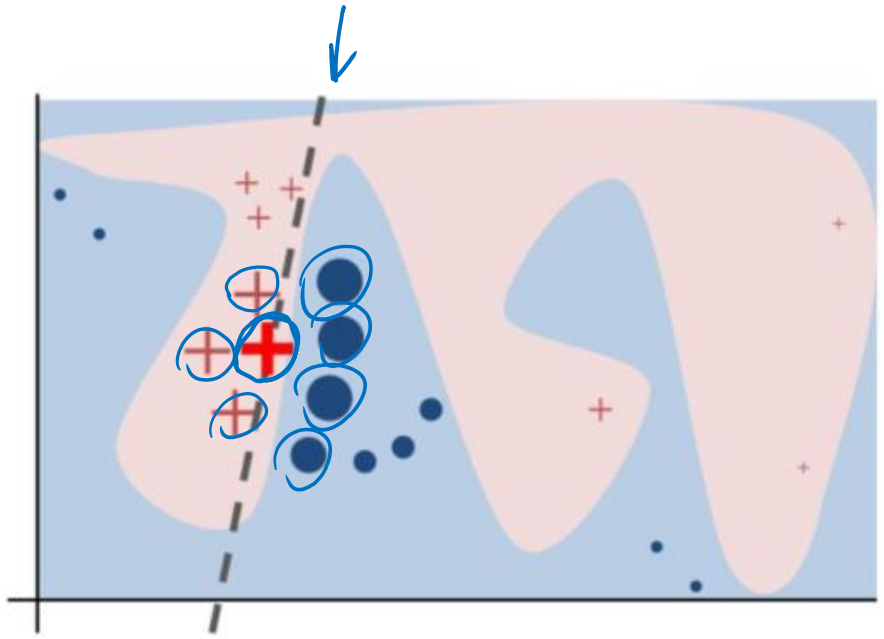
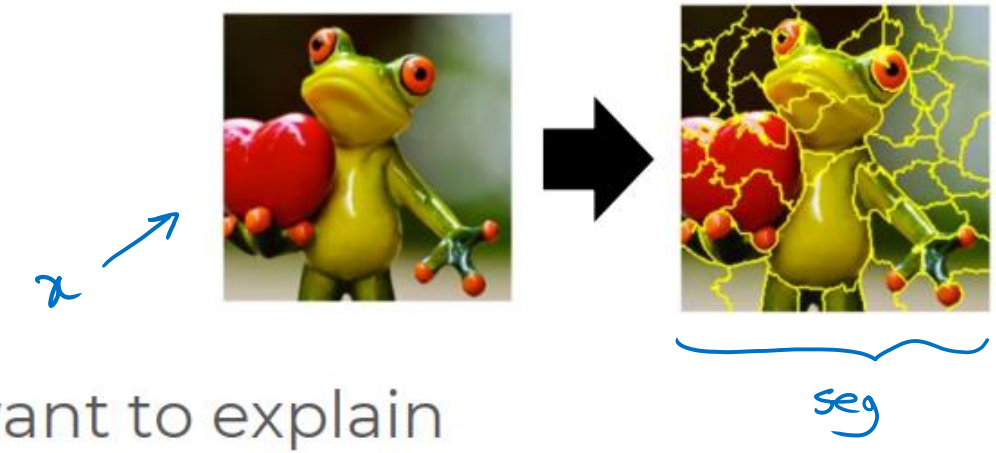


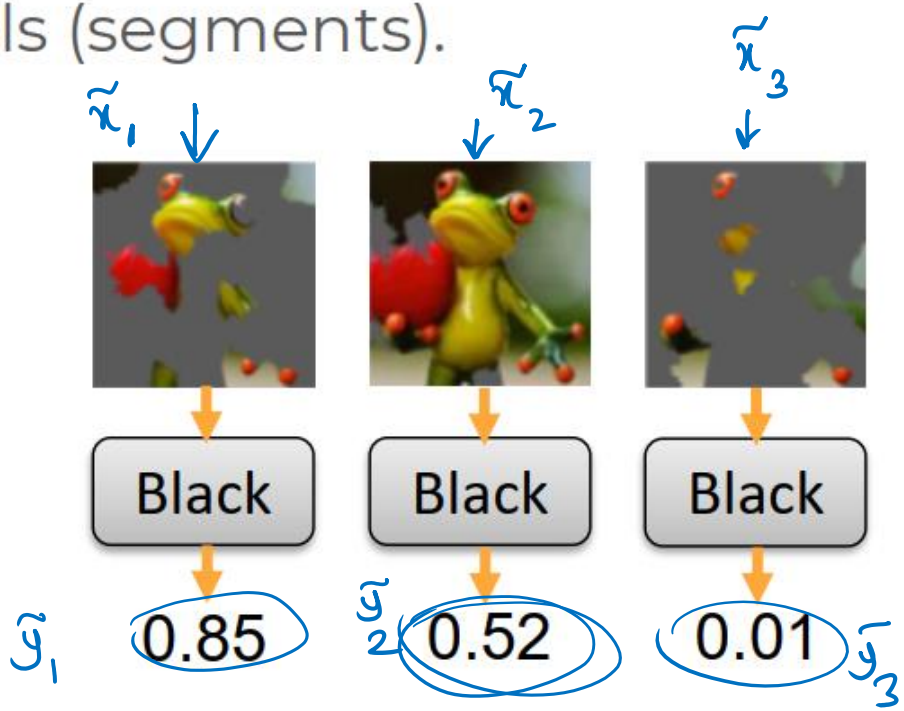
Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

[Ribeiro et al 2016]

LIME – Image



- 1. Given a data point you want to explain
- 2. Sample at the nearby - Each image is represented as a set of superpixels (segments).



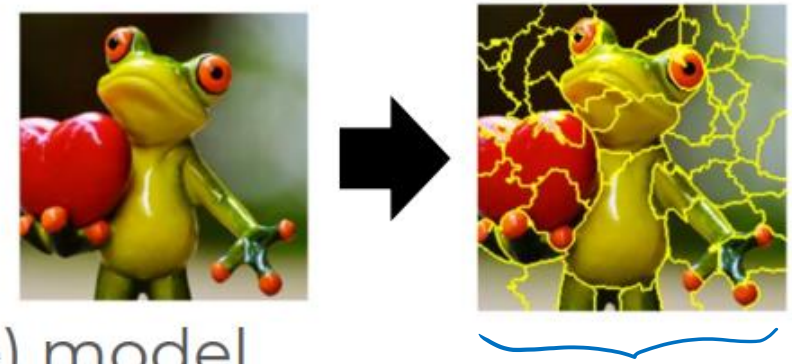
Randomly delete some segments.

Compute the probability of “frog” by black box

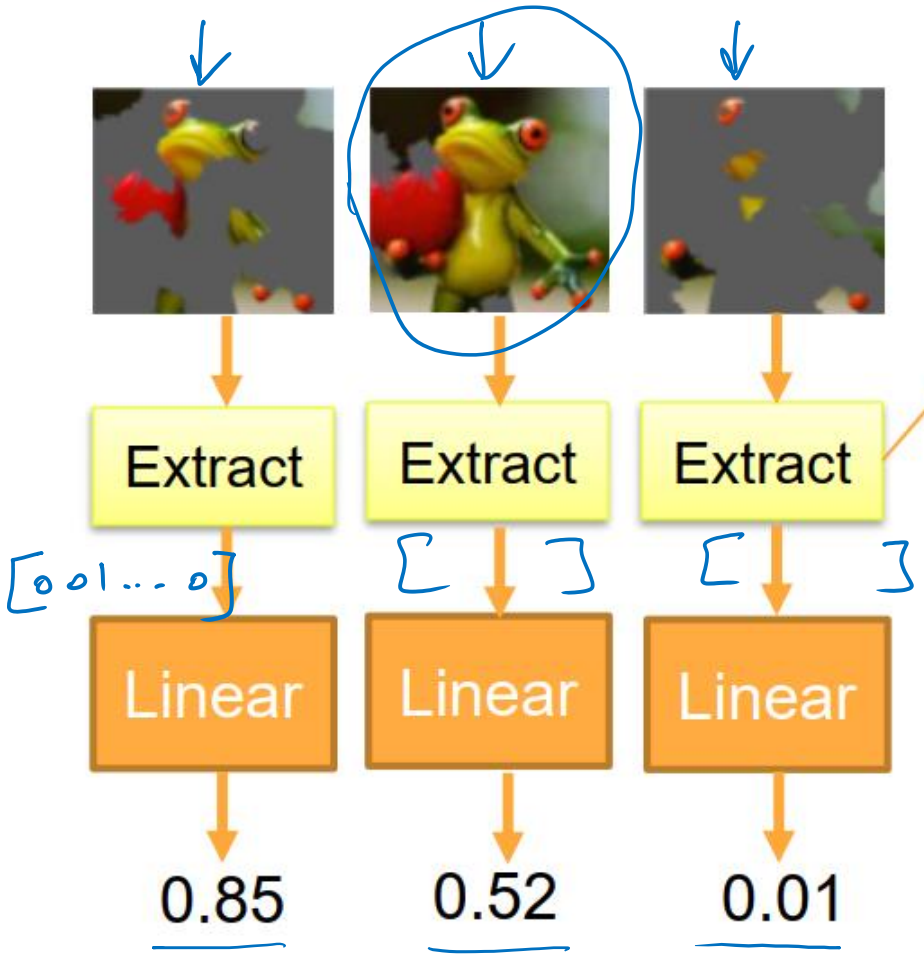
Ref: https://medium.com/@kstseng/lime-local-interpretable-model-agnostic-explanation%E6%8A%80%E8%A1%93%E4%BB%8B%E7%B4%B9-a67b6c34c3f8

(Slide Credit – Hung-yi Lee)

LIME – Image



- 3. Fit with linear (or interpretable) model



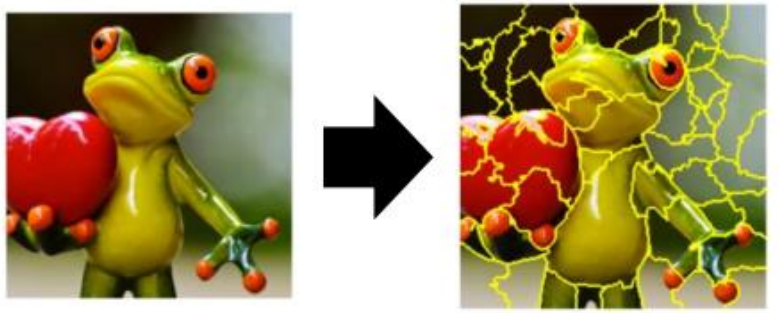
$x_1 \quad \cdots\cdots \quad x_m \quad \cdots\cdots \quad x_M$

$$x_m = \begin{cases} 0 & \text{Segment m is deleted.} \\ 1 & \text{Segment m exists.} \end{cases}$$

M is the number of segments.

(Slide Credit – Hung-yi Lee)

LIME – Image



- 4. Interpret the model you learned

Extract

Linear

0.85

$$y = w_1 x_1 + \cdots + w_m x_m + \cdots + w_M x_M$$

$$x_m = \begin{cases} 0 & \text{Segment m is deleted.} \\ 1 & \text{Segment m exists.} \end{cases}$$

M is the number of segments.

If $w_m \approx 0$ ➡ segment m is not related to “frog”

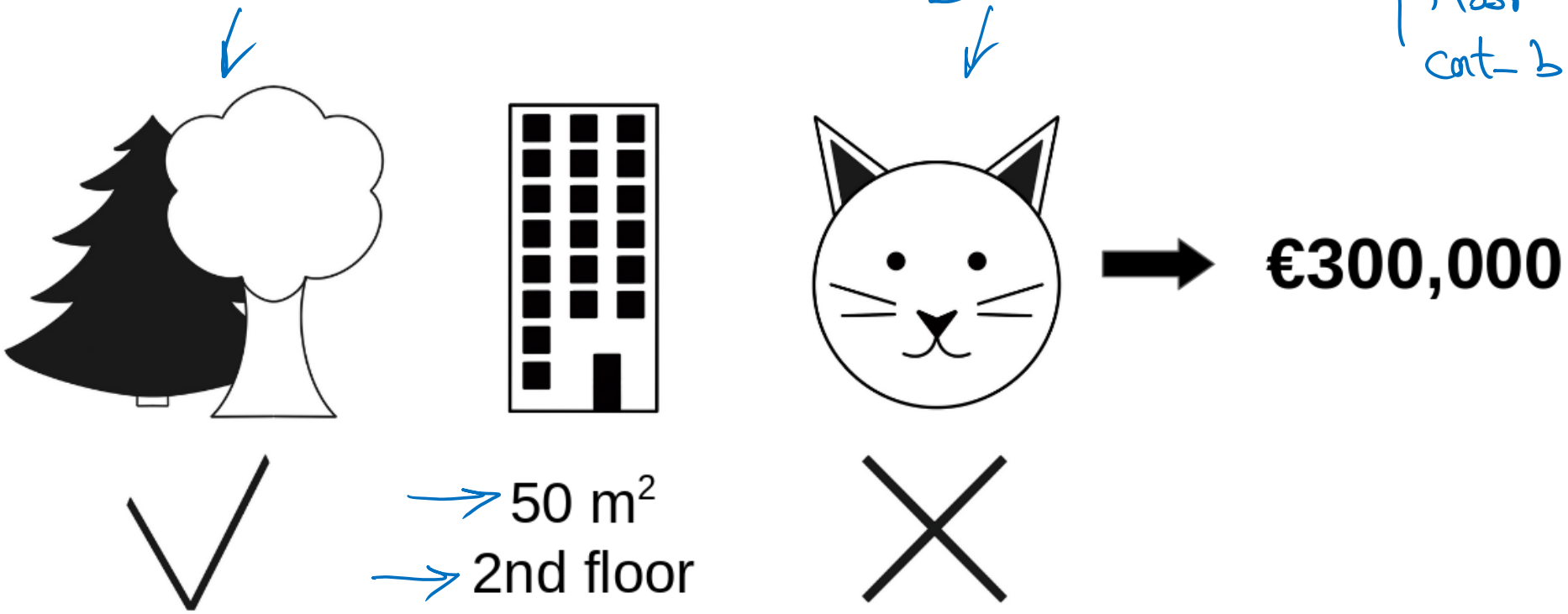
If w_m is positive ➡ segment m indicates the image is “frog”

If w_m is negative ➡ segment m indicates the image is not “frog”

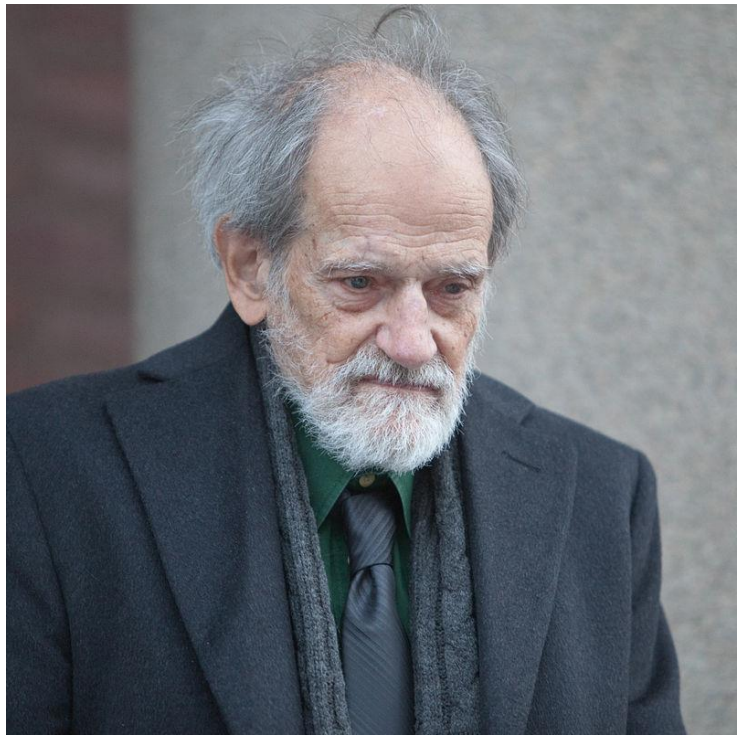
(Slide Credit – Hung-yi Lee)

Shapley values

- assuming that each feature value of the instance is a "player" in a game where the prediction is the payout.
- Example:
  - The average prediction for all apartments is €310,000



Shapley value



- The Shapley value is a solution concept in cooperative game theory.

- Lloyd Shapley introduced it in 1951 and won the Nobel Memorial Prize in Economic Sciences for it in 2012.

- How important is each player to the overall cooperation, and what payoff can he or she reasonably expect? The Shapley value provides one possible answer to this question.

Shapley value: Business Example

$v(\{0, w_1, w_2, w_3\}) = 3$

$v(\{0, w_1\}) = 1$

$v(\{0\}) = 0$

$v(\{w_1\}) = 0$

$0, w_1, w_2$

$v(\{0, w_1, w_2\}) = 2$

$1 \quad \frac{1}{2} \quad \frac{1}{2}$

$\frac{1}{3} \quad \frac{1}{3} \quad \frac{2}{3}$

Shapley value: Glove example

$N = \{1, 2, \dots, n\}$

Φ_i

coalition

$\{1,2,3\}$

L L R

$v(\{1,2,3\}) = 1$

$S \subset N$

$$v(S) = \begin{cases} 1 & \text{Se } \{\{1,3\},\{2,3\},\{1,2,3\}\} \\ 0 & \text{O.W.} \end{cases}$$

$v(S \cup \{i\}) - v(S)$

$v(\{1,2,3\}) = 10{,}000\ \$$

1	2	3
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
0	0	1
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{3}$

coalition	payoff
1	1000 $
1 2	6000
1 3	7000
1 2 3	10000

coalition	payoff
ϕ	0
2	5000
3	4000
2 3	9000

+1000
+1000
+3000
+1000

Shapley values

- Our goal is to explain the difference between the actual prediction (€300,000) and the average prediction (€310,000): a difference of -€10,000.

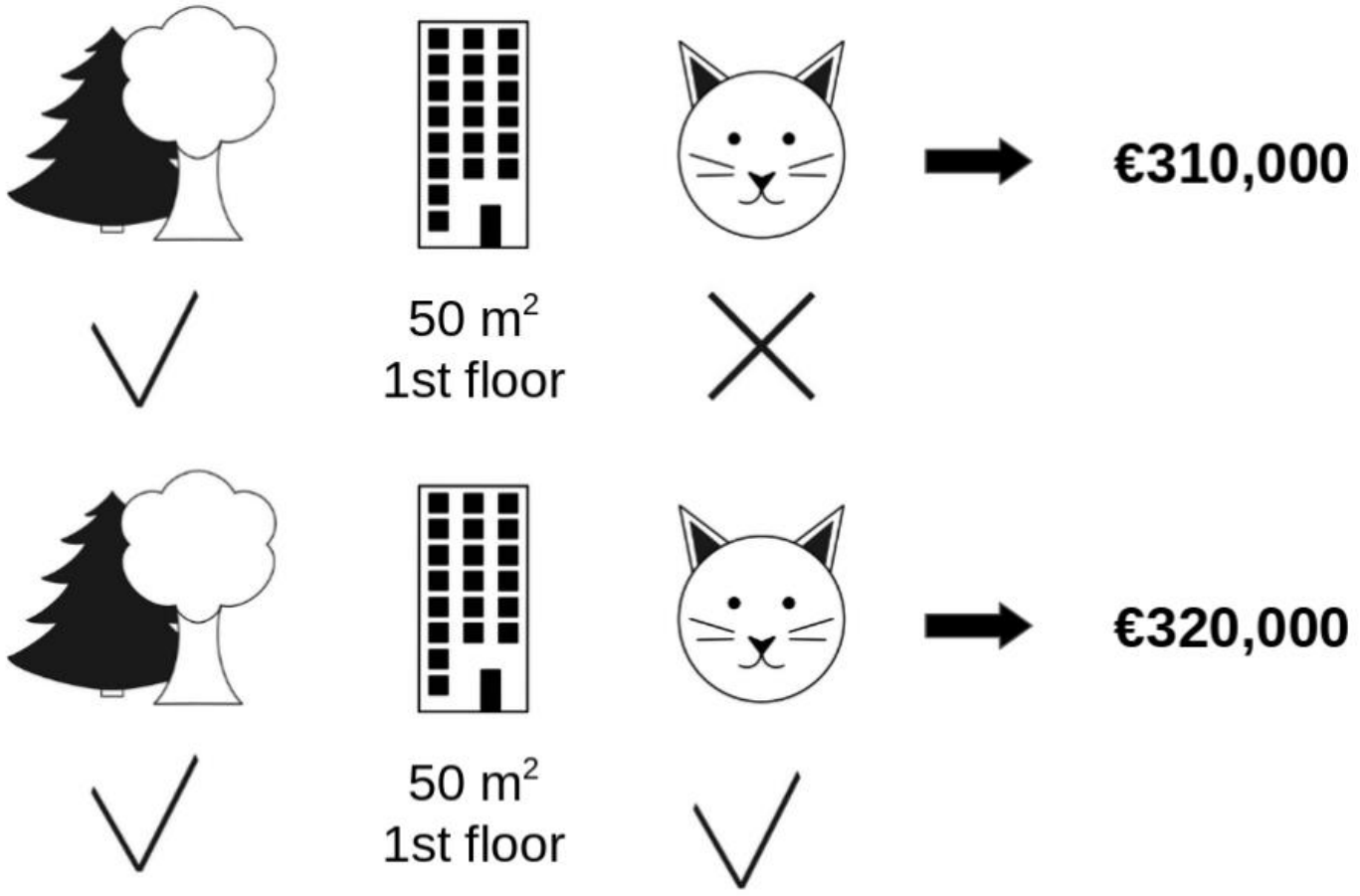
- The answer is simple for linear regression models. why?

- The answer could be:
  - park-nearby: €30,000
  - Area-50: €10,000
  - Floor-2nd: €0
  - cat-banned: -€50,000

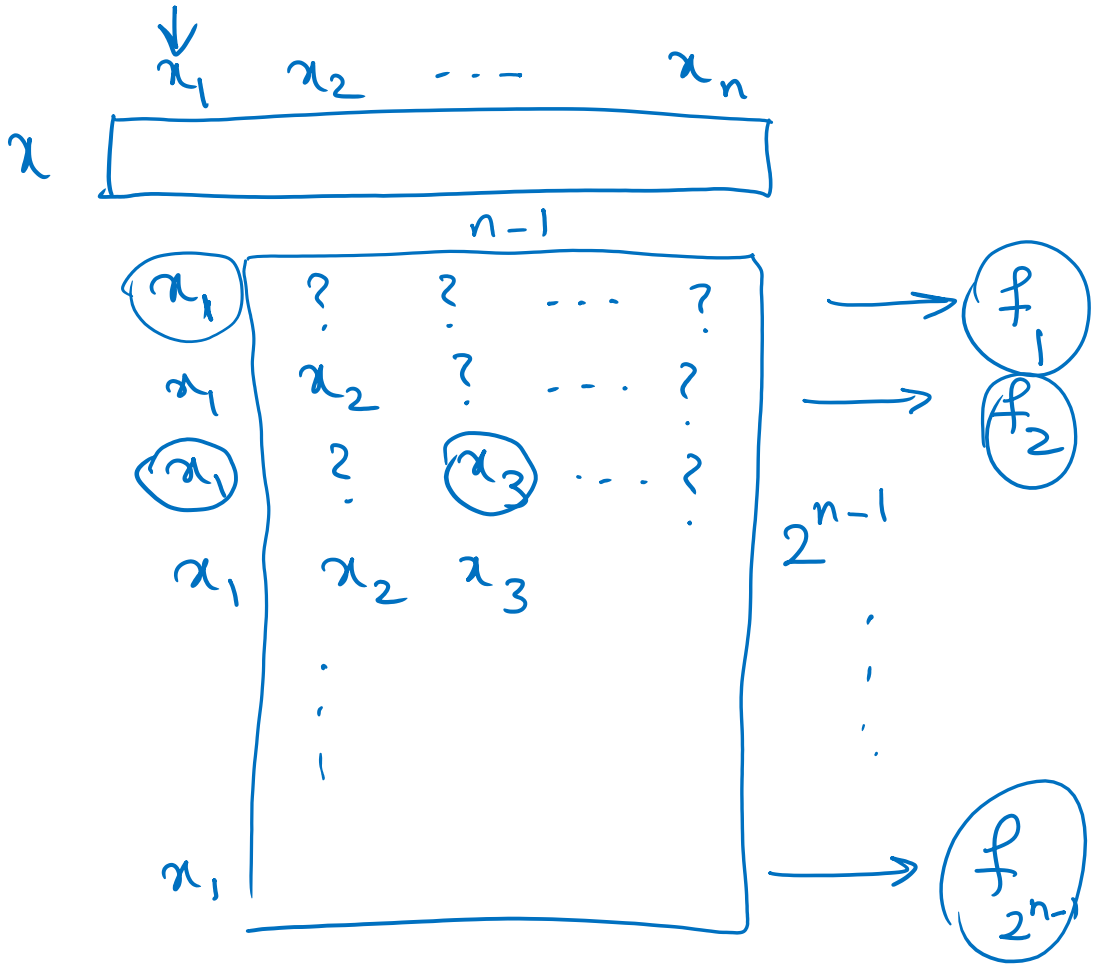
−10,000

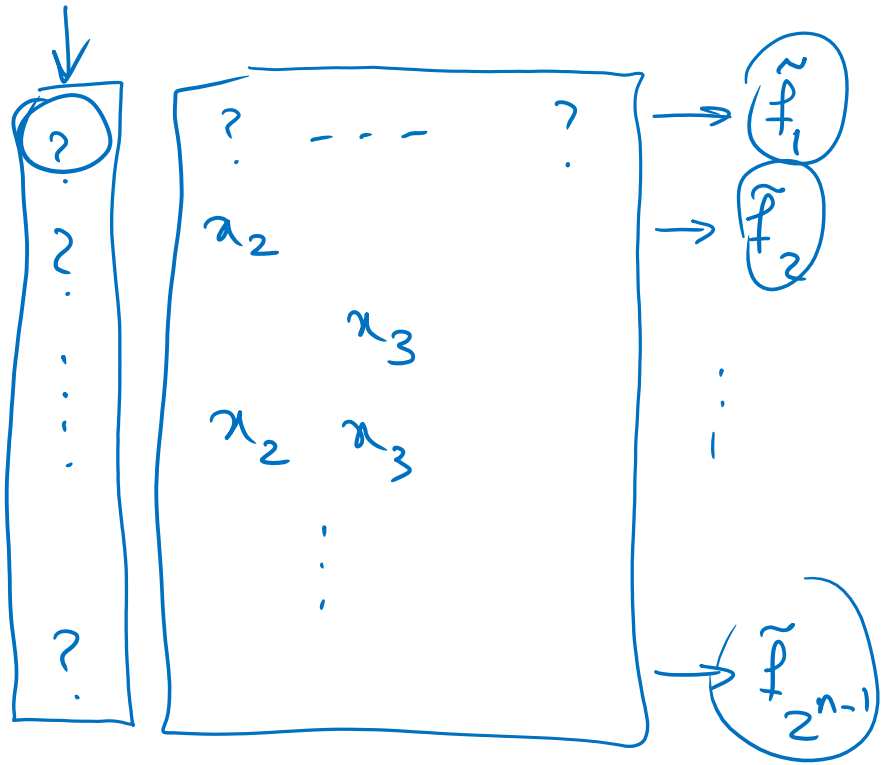
Shapley value

- The Shapley value is the average marginal contribution of a feature value across all possible coalitions.



observation

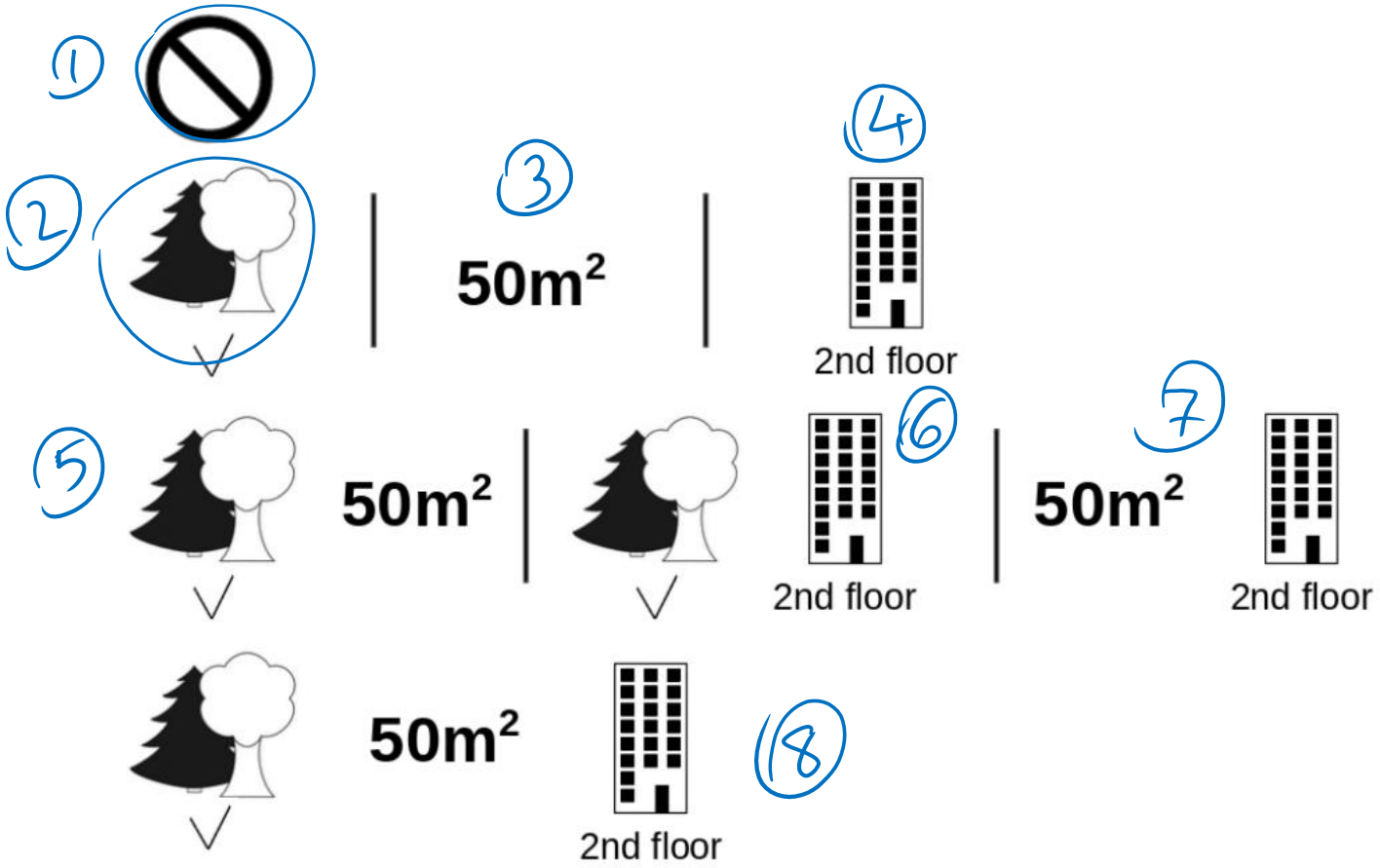




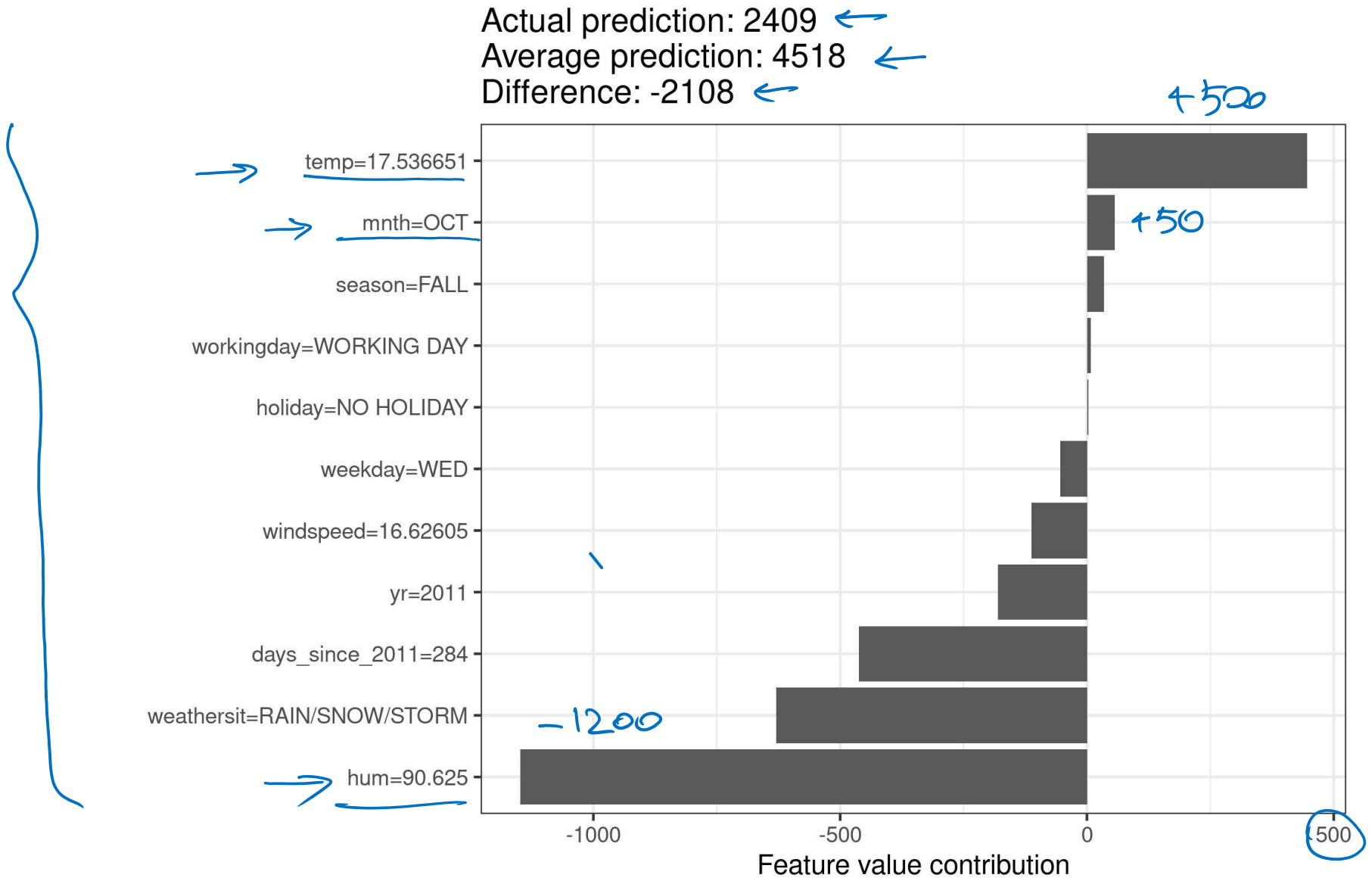
$O(n2^{n-1})$

Shapley value

- For each of these coalitions we compute the predicted apartment price with and without the feature value cat-banned and take the difference to get the marginal contribution. The Shapley value is the (weighted) average of marginal contributions.



Shapley value: Example



Shapley value advantages and disadvantages

- Advantages:
  - The difference between the prediction and the average prediction is fairly distributed among the feature values of the instance.
  - Solid theory

- Disadvantages:
  - High computation complexity
  - No prediction model.
  - You need access to the dataset
  - Inclusion of unrealistic data instances: dependent features