Shapley Analysis

Bertram Fuchs

bfuchs@student.ethz.ch

Spinal Cord Injury Artificial Intelligence Lab
ETH Zurich & Swiss Paraplegic Research

SSNR, Baiona | 20 June 2024

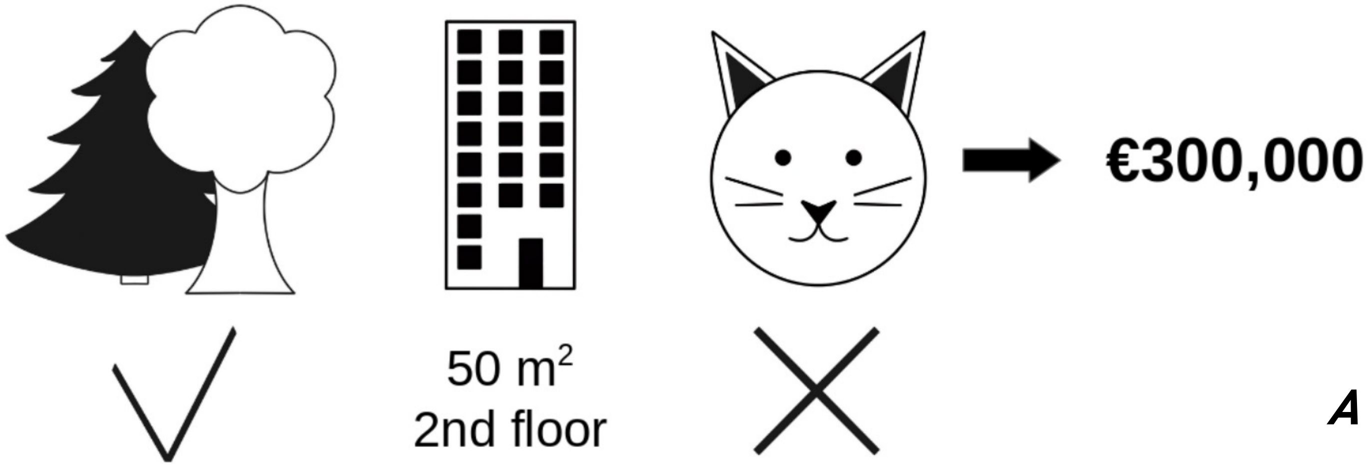
Shapley Values

Goal: Explain feature importances and individual feature values

Origin: Shapley values are a method from coalitional game theory.

“A prediction can be explained by assuming that each feature value of the instance is a “player” where the prediction is the payout.”

Example: Predict Apartment Prices | Average Prediction for all ins €310,000



Interpretable Machine Learning

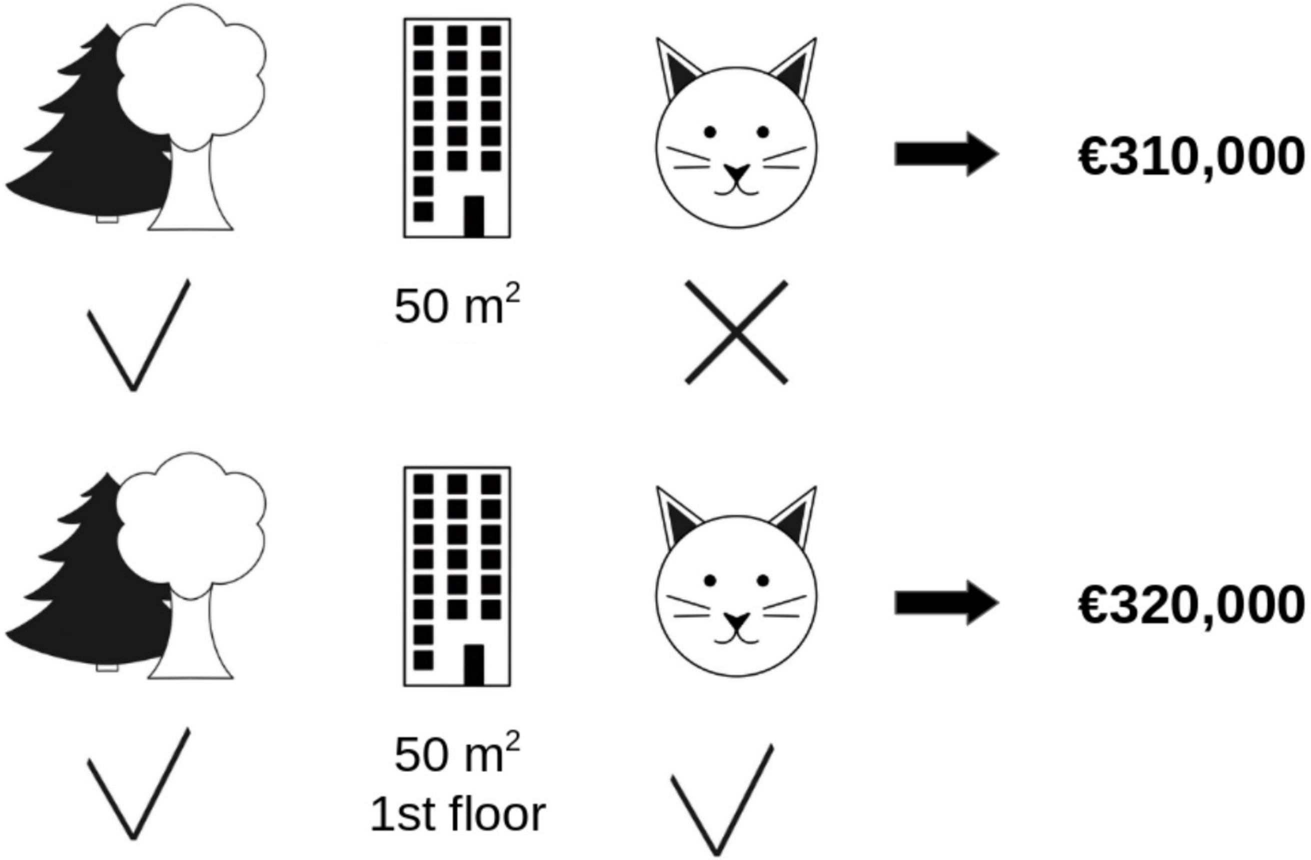
A Guide for Making Black Box Models Explainable

Christoph Molnar

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

Evaluate: Cat-banned feature

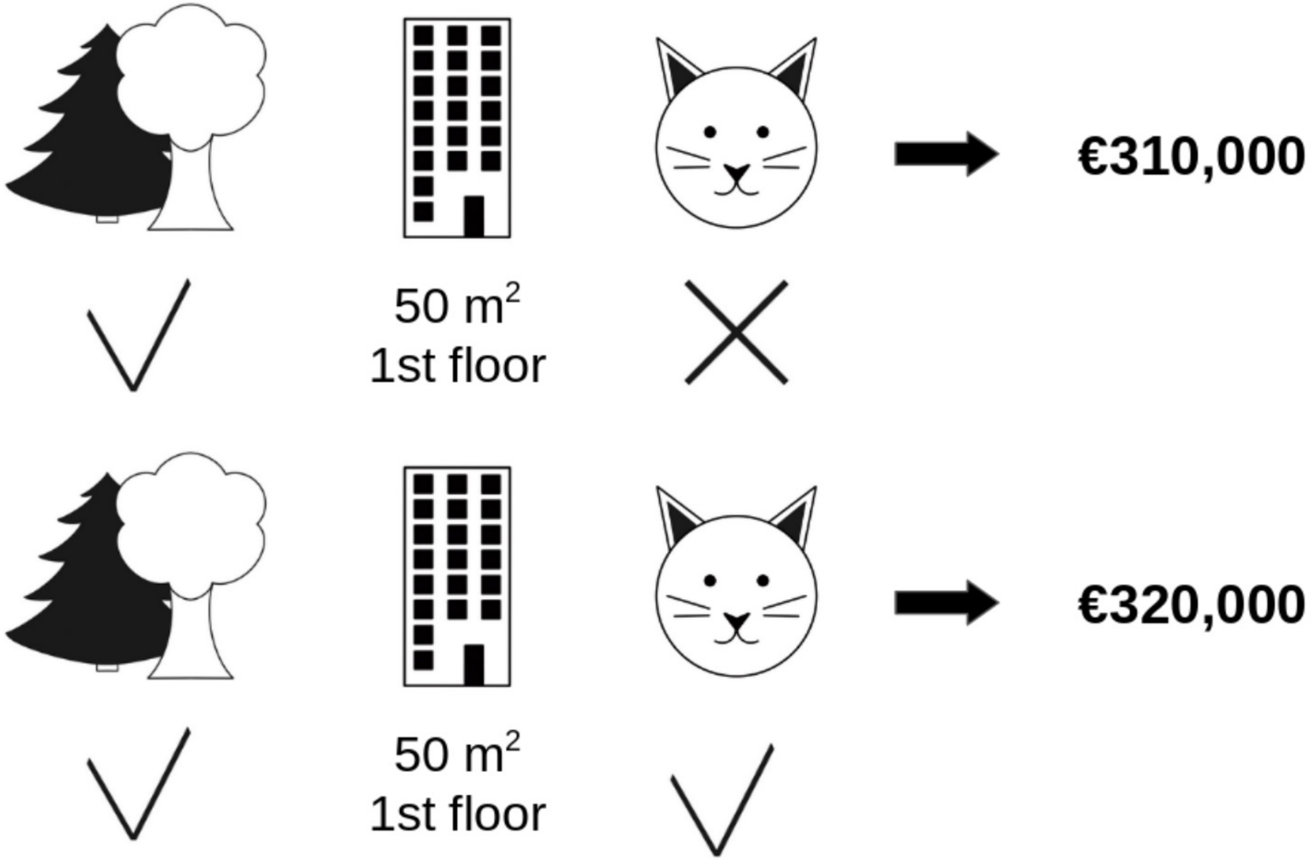
If feature not in coalition: randomly draw sample from data (1^{st} floor)



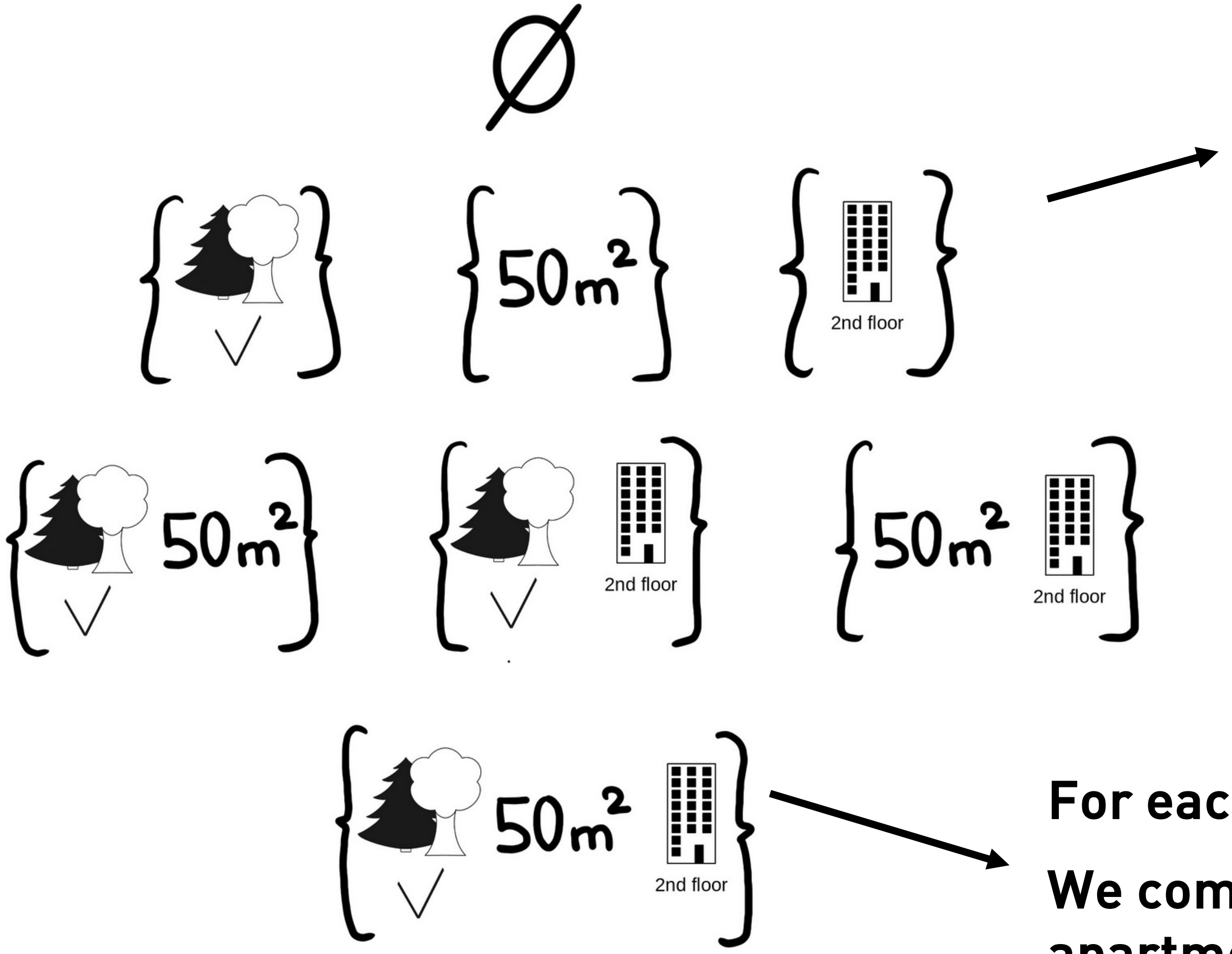
The Shapley value is the average marginal contribution of a feature value across all possible coalitions.

Evaluate: Cat-banned feature

If feature not in coalition:
randomly draw sample from data (1st floor)



We repeat this computation for all possible coalitions.



Feature values of features not in the coalition are filled with random feature values from the dataset.

For each coalition:

We compute the predicted apartment price with & w/o feature *cat-banned* and take the difference.

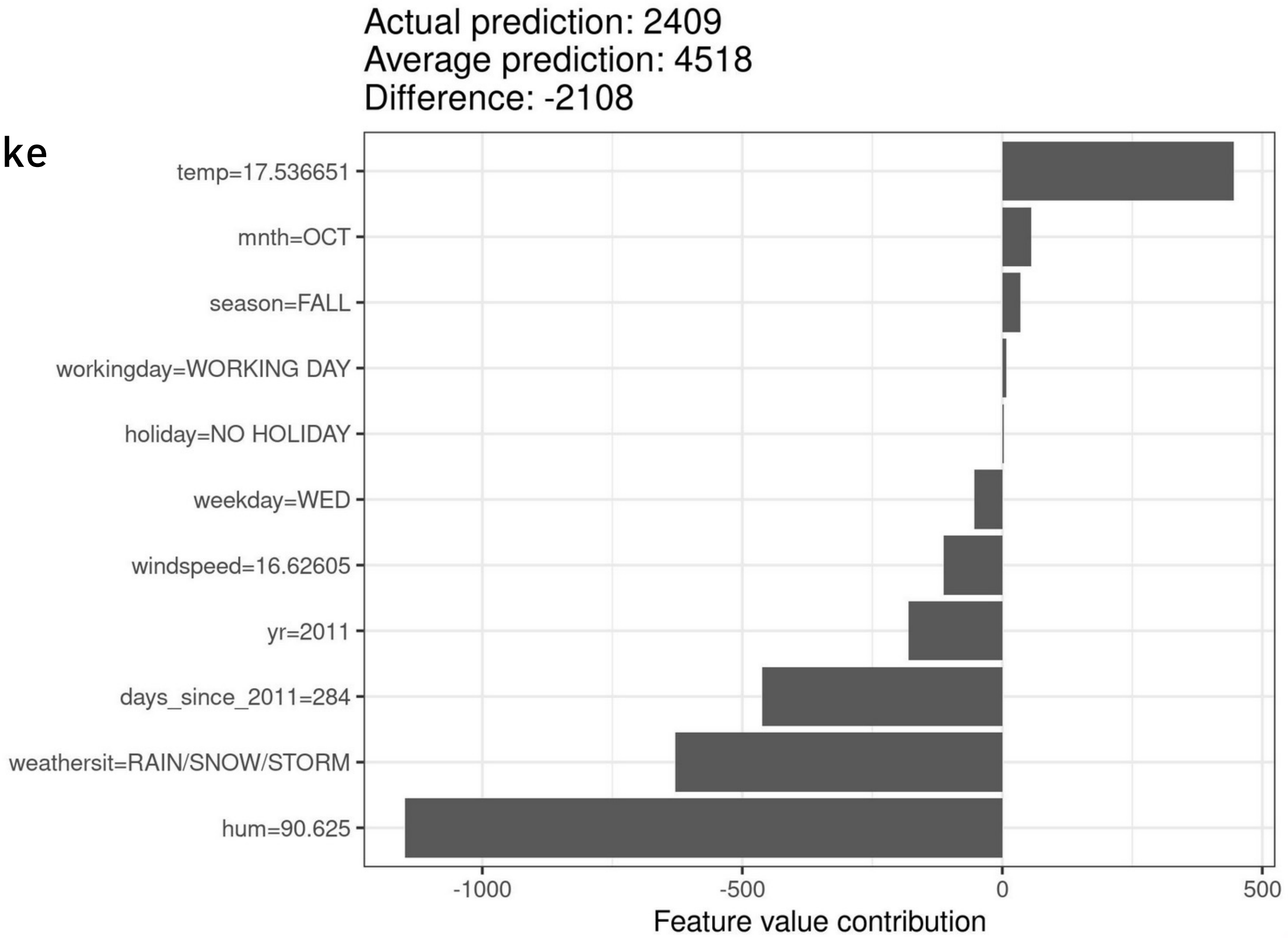
Example: Bike rental dataset (Random Forest model)

Prediction for one day: Day 285

Weater & humidity lead to a decrease in bike rentals

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{j\}\right) - val(S)\right)$$

$$val_x(S) = \int \hat{f}\,(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}\,(X))$$



TUTORIAL

Thank you!