XAI Techniques Cheatsheet

Yvo Keller

February 5, 2025

1 Tabular Methods

1.1 Shapley Values

1.1.1 Overview

Shapley values provide a way to fairly distribute the prediction among features by considering all possible feature combinations.

1.1.2 Key Formula

The Shapley value for feature i is:

$$\phi_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \times (|N| - |S| - 1)!}{|N|!} \left(f_\theta(S \cup \{i\}) - f_\theta(S)\right) \tag{1}$$

where:

- N is the set of all features
- S is a subset of features excluding feature i
- f_θ is the model prediction
- $|S|$ is the size of subset S
- $|N|$ is the total number of features

1.1.3 Calculation Process

1. Select an instance to explain
2. For each feature:
   (a) Generate all possible feature coalitions excluding the target feature
   (b) For each coalition:
      i. Calculate model prediction with and without target feature
      ii. Compute marginal contribution
      iii. Weight contribution by coalition size
   (c) Sum weighted contributions
3. Average contributions over all permutations

1.1.4 Properties

- **Efficiency**: Sum of Shapley values equals model output minus baseline
- **Symmetry**: Equal contribution features receive equal Shapley values
- **Dummy**: Features with no marginal contribution get zero Shapley value
- **Additivity**: Values can be computed independently and summed

1.1.5 Intuitive Example: Ice Cream Shop

Let's understand Shapley values through a practical example of predicting ice cream sales.

Setup Consider a model predicting daily ice cream sales with features:

- x_1 = Day of the week
- x_2 = Number of flights arriving
- x_3 = Temperature
- x_4 = Total opening hours

Calculation Process To calculate the Shapley value for temperature (x_3):

1. **Select a sample**: Choose a specific day's data point
2. **Choose baseline**: Select a reference point (usually average values)
3. **Generate permutation**: e.g., (x_4, x_1, x_3, x_2)
4. **Calculate marginal contributions**:
   - Start with baseline prediction: f_{base}
   - Add features one by one:

$$\begin{aligned}&f(x_4)\\&f(x_4, x_1)\\&f(x_4, x_1, x_3) \leftarrow \text{Temperature added here}\\&f(x_4, x_1, x_3, x_2)\end{aligned}$$

   - Temperature's contribution $= f(x_4, x_1, x_3) - f(x_4, x_1)$
5. **Repeat**: Do this for multiple permutations
6. **Average**: The Shapley value is the average contribution across permutations

Interpretation The final Shapley value for temperature tells us:

- Positive value: Higher temperatures increase ice cream sales
- Negative value: Higher temperatures decrease sales
- Magnitude: Size of temperature's impact on the prediction

1.1.6 Monte Carlo Approximation

For large feature sets, exact computation becomes infeasible. Monte Carlo approximation:

1. Sample random feature permutations
2. Calculate marginal contributions for each permutation
3. Average results over all samples

1.2 Partial Dependence Plots (PDP)

1.2.1 Overview

PDPs show how a feature affects predictions on average, while marginalizing over all other features.

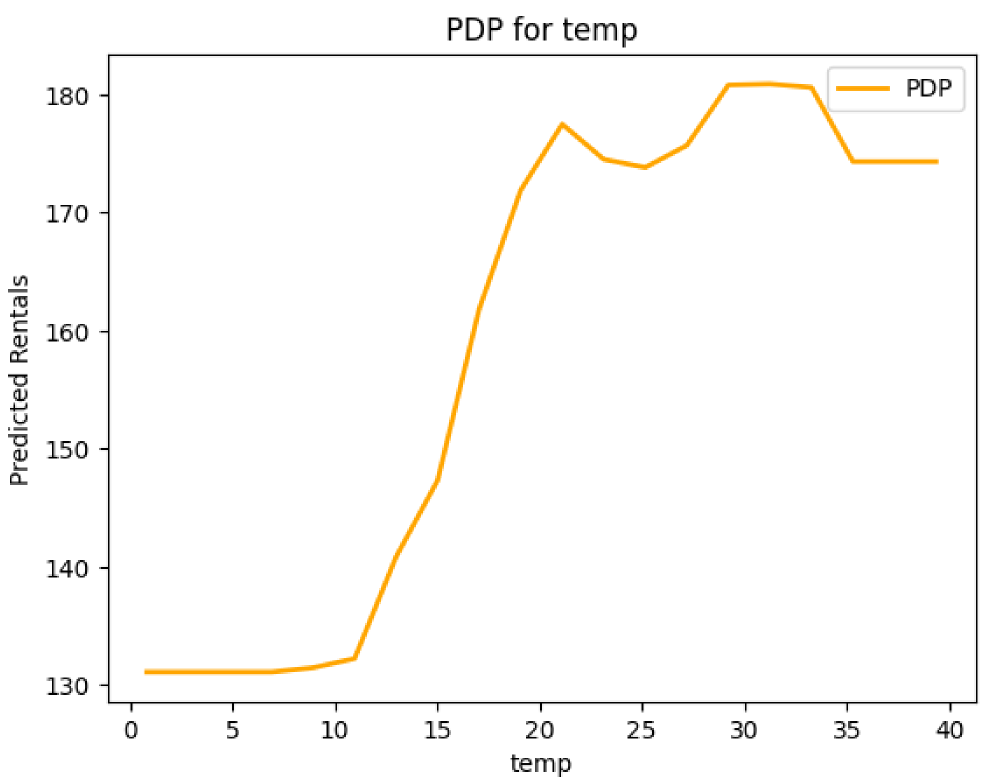

Figure 1: Example of a Partial Dependence Plot (PDP)

1.2.2 Intuitive Example

Consider our ice cream sales model:

- To create a PDP for temperature:
  1. Pick a temperature value (e.g., 25°C)
  2. For every data point, set temperature to 25°C
  3. Get model predictions for all these modified points
  4. Average these predictions
  5. Repeat for different temperature values
  6. Plot temperature vs. average predictions

1.2.3 Interpretation

- Slope shows relationship strength
- Shape reveals non-linear effects
- Flat regions indicate no impact
- Limitations: Can miss feature interactions

1.3 Individual Conditional Expectation (ICE)

1.3.1 Overview

ICE plots extend PDPs by showing how predictions change for individual instances, revealing heterogeneous effects hidden by PDPs.

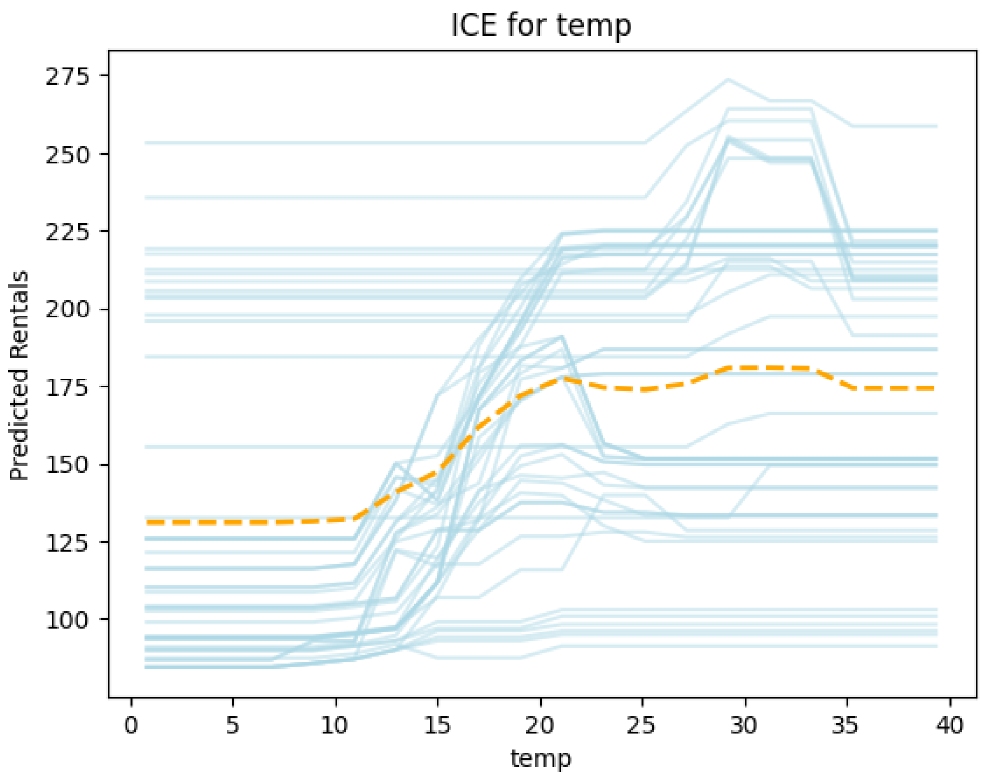

Figure 2: Example of Individual Conditional Expectation (ICE) plots

1.3.2 Intuitive Example

Using our ice cream model:

- For each individual day in our dataset:
  1. Keep all features fixed except temperature
  2. Vary temperature across its range
  3. Plot prediction line for this specific day
- Result: Multiple lines, one per instance
- PDP would be the average of these lines

1.3.3 Key Insights

- Diverging lines suggest feature interactions
- Parallel lines indicate consistent effects
- Crossing lines show complex relationships
- More informative than PDP alone

1.3.4 When to Use

- Feature interaction analysis
- Detecting heterogeneous effects
- Model behavior validation
- Identifying outlier instances