Shapley values and SHAP

Based on 'Interpretable Machine Learning: A Guide for Making Black Box Models Explainable' and the free sample 'Interpreting Machine Learning Models With SHAP' by Christoph Molnar

Dawid Kubkowski

May 8, 2025

Motivation

- Three friends share a taxi ride home: Alice, Bob, and Charlie.
- Total fare: \$51.
- How to split costs fairly?
- This simple problem introduces the idea of fair cost allocation, which is central to Shapley values.
- We aim to translate this concept to interpreting machine learning models.

Coalitional Game Basics

- A *coalitional game* is a pair (N, v) where:
  - N is the set of players (here $\{A, B, C\}$).
  - $v(S)$ is the value (or cost) of any subset $S \subseteq N$.
- Here we use cost function $c(S) = v(S)$
- We seek an allocation (x_A, x_B, x_C) such that $x_A + x_B + x_C = c(N)$.

Cost Function $c(S)$

Passengers	**Cost**	**Note**
$\emptyset$	\$0	No taxi, no cost
{Alice}	\$15	Standard fare to Alice's and Bob's place
{Bob}	\$25	Bob insists on luxury taxi
{Charlie}	\$38	Charlie lives farther away
{Alice, Bob}	\$25	Bob always gets his way
{Alice, Charlie}	\$41	Drop off Alice then Charlie
{Bob, Charlie}	\$51	Drop off luxurious Bob first, then Charlie
{Alice, Bob, Charlie}	\$51	All three together

Marginal contribution

$$\Big(c(S \cup \{i\}) - c(S)\Big)$$

The marginal contribution of a player to a coalition is the value of the coalition with the player minus the value of the coalition without the player.

Addition to Coalition and Marginal Contributions

Player	Added to	Cost Before	Cost After	Marginal Contribution
Alice	$\emptyset$	\$0	\$15	\$15
Alice	{Bob}	\$25	\$25	\$0
Alice	{Charlie}	\$38	\$41	\$3
Alice	{Bob, Charlie}	\$51	\$51	\$0
Bob	$\emptyset$	\$0	\$25	\$25
Bob	{Alice}	\$15	\$25	\$10
Bob	{Charlie}	\$38	\$51	\$13
Bob	{Alice, Charlie}	\$41	\$51	\$10
Charlie	$\emptyset$	\$0	\$38	\$38
Charlie	{Alice}	\$15	\$41	\$26
Charlie	{Bob}	\$25	\$51	\$26
Charlie	{Alice, Bob}	\$25	\$51	\$26

Permutations of passengers

Simply averaging these contributions treats all situations equally, but some scenarios—like adding a person to an empty taxi—provide more insight into fair cost distribution. To better assess each passenger's impact, one method is to examine all possible orderings (permutations) of the passengers.
There are $3! = 6$ possible orderings:

- Alice, Bob, Charlie
- Alice, Charlie, Bob
- Bob, Alice, Charlie
- Charlie, Alice, Bob
- Bob, Charlie, Alice
- Charlie, Bob, Alice

Averaging marginal contributions

In two of these cases, Alice was added to an empty taxi, and in one case, she was added to a taxi with only Bob. By weighting the marginal contributions accordingly, we calculate the following weighted average marginal contribution for Alice, abbreviating Alice, Bob, and Charlie to A, B, and C:

$$\frac{1}{6}\left(\underbrace{2\cdot\$15}_{\text{A to }\emptyset}+\underbrace{1\cdot\$0}_{\text{A to B}}+\underbrace{1\cdot\$3}_{\text{A to C}}+\underbrace{2\cdot\$0}_{\text{A to B,C}}\right)=\$5.50$$

We multiply by $\frac{1}{6}$ because 6 is the sum of the weights $(2+1+1+2)$. That's how much Alice should pay for the ride: \$5.50.

Alice: \$5.50
Bob: \$15.50
Charlie: \$30.00
Sum: \$51.00

Shapley Value Allocation

- Shapley value allocates cost by average marginal contributions:

$$\phi_i(c) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Big(c(S \cup \{i\}) - c(S) \Big).$$

- Compute each player's contributions:
  - ϕ_A: ... (detailed calculation)
  - ϕ_B: ...
  - ϕ_C: ...
- These sum to total cost 51.

Axioms behind Shapley values

- **Efficiency**:

$$\sum_i \phi_i(v) \;=\; v(N)$$

  The total payoff is exactly the worth of the grand coalition.
- **Symmetry**: If players i and j are interchangeable (i.e., for all coalitions $S \subseteq N \setminus \{i,j\}$, $v(S \cup \{i\}) = v(S \cup \{j\})$), then $\phi_i(v) = \phi_j(v)$.
- **Null Player**: If player i contributes no marginal value to any coalition ($v(S \cup \{i\}) = v(S)$ for all S), then $\phi_i(v) = 0$.
- **Additivity**: For any two games v and w, the Shapley value satisfies

$$\phi(v + w) \;=\; \phi(v) \;+\; \phi(w).$$

Machine Learning example

Consider the following scenario:
You have trained a machine learning model f to predict apartment prices.
For a specific apartment $x^{(i)}$, the model predicts:

$$f(x^{(i)}) = 300{,}000$$

Your task is to explain this prediction. The apartment $x^{(i)}$ has the following features:

- Area: 50 m^2 (538 square feet)
- Floor: 2nd
- Nearby park: Yes
- Cats: Banned

These features are what the model used to make the prediction. The average prediction for all apartments in the data is €310,000, which places the predicted price of this specific apartment slightly below average.

Explaining the Prediction

Goal: Explain the difference between the actual prediction (€300,000) and the average prediction (€310,000).

Difference: $-10{,}000$

Feature Contributions:

- `area-50`: $+10{,}000$
- `floor-2nd`: 0
- `park-nearby`: $+30{,}000$
- `cat-banned`: $-50{,}000$

Total Contribution: $-10{,}000 = 300{,}000 - 310{,}000$

Concepts and Corresponding ML Terms

Concept	**Machine Learning**	**Term**
Player	Feature index	j
Coalition	Set of features	$S \subseteq \{1, \ldots, p\}$
Not in coalition	Features not in the coalition	$C = \{1, \ldots, p\} \setminus S$
Coalition size	Number of features in the coalition	$\lvert S \rvert$
Total players	Total number of features	p
Total payout	Prediction for $x^{(i)}$ minus the average prediction	$f(x^{(i)}) - \mathbb{E}(f(X))$
Value function	Prediction for values in coalition S minus expected	$v_{f,x^{(i)}}(S)$
SHAP value	Contribution of feature j to the total payout	$\phi_j^{(i)}$

The SHAP value function

The SHAP value function, for a given model f and data instance $x^{(i)}$, is defined as:

$$v_{f,x^{(i)}}(S) = \int f(x_S^{(i)} \cup X_C)\, d\mathbb{P}_{X_C} - \mathbb{E}[f(X)]$$

$x_S^{(i)} \cup X_C$ is a feature vector $\in \mathbb{R}^p$ where values at positions S have values from $x_S^{(i)}$ and the rest are random variables from X_C.

Value function: apartment example

Park	Cat	Area	Floor	Predicted Price
Nearby	Banned	50	2nd	€300,000

Informally, the value function for the coalition of `park` and `floor` would be:

$$v(\{\text{park}, \text{floor}\}) = \int f(x_{\text{park}}, X_{\text{cat}}, X_{\text{area}}, x_{\text{floor}})\, d\mathbb{P}_{X_{\text{cat}}, X_{\text{area}}} - \mathbb{E}_X(f(X))$$

where $x_{\text{park}} = \text{nearby}$, $x_{\text{floor}} = 2$, and $\mathbb{E}_X(f(X)) = 300{,}000$.

- The features `park` and `floor` are *present*, so we input their corresponding values into f.
- The features `cat` and `area` are *absent*, and thus are treated as random variables and integrated over.

SHAP Value Equation

Combining all the terms into the Shapley value equation, we get the SHAP equation:

$$\phi_j^{(i)} = \sum_{S \subseteq \{1,\dots,p\} \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} \cdot \left(\int f\left(x^{(i)}_{S \cup \{j\}} \cup X_{C \setminus \{j\}}\right) \, d\mathbb{P}_{X_{C \setminus \{j\}}} - \int f\left(x^{(i)}_S \cup X_C\right) \, d\mathbb{P}_{X_C} \right)$$

The SHAP value $\phi_j^{(i)}$ of a feature value is the average marginal contribution of a feature value $x_j^{(i)}$ to all possible coalitions of features.

Estimating Shapley Values

Exact vs Approximate

- *Exact*: Sum over all 2^k coalitions—impractical for large k.
- *Approximate*: Monte Carlo sampling (Štrumbelj Kononenko, 2014).

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^{M} \left(f(x_{+j}^{(m)}) - f(x_{-j}^{(m)}) \right)$$

Key Definitions

- $f(x_{+j}^{(m)})$: prediction for x with feature j fixed, others replaced by values from a random draw.
- $f(x_{-j}^{(m)})$: prediction with feature j replaced by random draw, others fixed.
- Each of the M "Frankenstein" instances mixes x with a random reference $\mathbf{z}$.

Pros and cons

Strengths

- **Efficiency**: Fair distribution of the prediction difference across features (efficiency axiom).
- **Contrastive explanations**: Can compare to a subset of data or even a single reference instance.
- **Theoretical foundation**: Axioms (efficiency, symmetry, dummy, additivity) ensure consistency and justification.

Limitations

- **Computational cost**: Exponential number of coalitions forces approximate sampling.
- **Sampling variance**: No clear guideline for choosing the number of iterations M.
- **Interpretation pitfalls**:
  - Not a local derivative—does not imply gradient-like behavior.
  - A positive Shapley value does *not* guarantee that increasing the feature increases the prediction.

Acknowledgements

- This presentation is based on the book:
  Interpreting Machine Learning Models With SHAP by Christoph Molnar,
  Interpretable Machine Learning: A Guide for Making Black Box Models Explainable by Christoph Molnar.
- Free sample of *Interpreting Machine Learning Models With SHAP*:
  `https://leanpub.com/shap`
- Full online version of the book:
  `https://christophm.github.io/interpretable-ml-book`