$f(x) = w_0 + \overline{w_1 x_1} + \overline{w_2 x_2} + \overline{w_3 x_3}$

$w_2(x_2 - x_2')$

Effect

$w_2 x_2$

Additive Feature Attribution

multicollinearity

$x \xrightarrow{h_x} z$ binary

$f(x)$ original model

$g(z)$

Additive Feature Attribution Methods (AFAM)

- Additive feature attribution methods have an explanation model that is a linear function of binary variables

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i'$$

where $z' \in \{0,1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$

- ϕ_i is the effect attributed to the i-th feature.
- Summing the effects of all feature attributions approximates $f(x)$

$g(z') \simeq f(x)$

Simple properties uniquely determine AFA

- The AFA Methods will admit a unique solution with three desirable properties
  - Local Accuracy
  - Missingness
  - Consistency

ϕ_i

AFAM solution property — Local Accuracy

- Original model f
- The local accuracy requires the explanation model to at least match the output of f for the simplified input x' (which corresponds to the original input x)

$x \to x'$

$h_x(x')$

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i'$$

The explanation model g(x′) matches $f(x)$ when $x = h_x(x')$

f $\quad z'$ $\quad z' \setminus i$

AFAM solution property — Missingness

- If the simplified inputs represent feature presence, then missingness requires features missing in the original input to have no impact.

$$x_i' = 0 \implies \phi_i = 0$$

- Features which are not present will have no attributed impact

AFAM solution property — Consistency

- If one model gives a higher prediction for a particular feature then that model will have a higher ϕ_i for that feature.
- Let $f_x(z') = f(h_x(z'))$ and $z' \setminus i$ denotes setting $z_i = 0$

$f \quad f'$

- For any two models f and f',

if $\quad f_x'(z') - f_x'(z_i \setminus i) \geq f_x(z') - f_x(z' \setminus i)$ for all inputs $z' \in \{0,1\}^M$,

then $\quad \phi_i(f', x) \geq \phi_i(f, x)$

$h_x(\boxed{z'}) \to f \to f(h_x(z'))$ $\quad f_x(z')$

simplified

• z'

$\underbrace{f_x \quad f_x'}_{\text{2 such functions}}$

A result from cooperative game theory

- There is only one possible explanation model g which is linear (AFAM) and satisfies the three properties.

→ x original
x' simplified
z' binary vector

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \backslash i)\right]$$

different coalitions

x_j x'_j

x'	z'
non zero	1
zero	0

$|z'|$ is the number of non-zero entries in z'
$z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x'

- ϕ_i are the shapely values

Original x → Simplified space x'

Shapely Regression Values

- Shapely regression values are feature importances for linear models in the presence of multicollinearity
- This method requires retraining the model on all feature subsets $S \subseteq F$ where F is the set of all features
- It assigns an importance value to each feature that represents the effect on the model prediction of including that feature.

$f(\text{subset})$

Estimating Shapely Regression Values

$f_{S \cup \{i\}}$

- A model $f_{S \cup \{i\}}$ is trained with the i-th feature present
- Another model f_S is trained with the i-th feature withheld
- The predictions are compared on the current input

f_S

$$f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$$

$x_S \ i / i$

- The differences are computed for all possible subsets $S \subseteq F \backslash \{i\}$
- The effect of withholding a feature depends on other features in the model.

- The Shapely values are the weighted average of all possible differences:

$$\phi_i = \sum_{S \subseteq F \backslash \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)\right]$$

- The Shapely values are used as feature attributions

- For AFAM equivalence
Consider simplified binary inputs are mapped to the original feature space as —
1: original input
0: exclusion from the model

Shapely Sampling Values

- The shapely sampling values can be used to explain *any* model
- The effect of removing a variable from the model is approximated by integrating over samples from the training dataset.

→ $f(x_S)$ selected features kept as is.

$f(x_S \backslash i)$ other features sampled