$$\psi_{1}(x_{2}-x_{1}^{2}) = \psi_{0} + \overline{\psi_{1}}x_{1} + \overline{\psi_{2}}x_{2} + \overline{\psi_{3}}x_{3}$$

$$\psi_{1}(x_{2}-x_{2}^{2}) = \psi_{0} + \overline{\psi_{1}}x_{1} + \overline{\psi_{2}}x_{2} + \overline{\psi_{3}}x_{3}$$

Additive Feature Attribution



Additive Feature Attribution Methods (AFAM)

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

$$\underline{g(z')} = \underbrace{\phi_0}_{i=1} + \sum_{i=1}^{M} \underbrace{\phi_i z'_i}_{i}$$

 $\underline{g(z')} = \underbrace{\phi_0}_{i=1} + \sum_{i=1}^M \underbrace{\phi_i z_i'}_{i}$ where $\underline{z'} \in \{0,1\}^M$, 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)

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

Simple properties uniquely determine AFA

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



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^\prime (which corresponds to the original input x)

$$\underline{f(x)}=\underline{g(x')}=\phi_0+\sum_{i=1}^M\phi_ix_i'$$
 The explanation model $\underline{g(x')}$ matches $f(x)$ when $x=h_x(x')$

AFAM solution property — Missingness

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

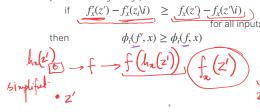
$$x_i' = 0 \implies \phi_i = 0$$

· Features which are not present will have no attributed impact

AFAM solution property — Consistency

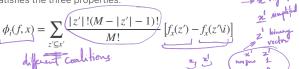
- If one model gives a higher prediction for a particular feature then that model will have a higher ϕ_i for that feature.
- Let $f_{\mathbf{x}}(z') = f(h_{\mathbf{x}}(z'))$ and $z' \setminus i$ denotes setting $z_i = 0$
- For any two models f and f',

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



A result from cooperative game theory

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



|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^\prime

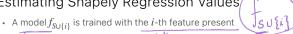
• ϕ_i are the shapely values

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.



Estimating Shapely Regression Values



- Another model f_S is trained with the feature withheld
- The predictions are compared on the current input



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

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

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

- The Shapely values are used as feature attributions
- For AFAM equivalence Consider simplified binary inputs are mapped to the original feature
 - 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.

